Proceedings of the
OAGM Workshop 2021

Computer Vision and Pattern Analysis Across Domains

November 24-25, 2021
University of Applied Sciences
St. Pölten, Austria

OAGM - Austrian Association for Pattern Recognition

TU Graz
Graz University of Technology

OAGM

OESTERREICHISCHE COMPUTER GESELLSCHAFT
AUSTRIAN COMPUTER SOCIETY

Markus Seidl, Matthias Zeppelzauer, and Peter M. Roth (eds.)

Proceedings of the

OAGM Workshop 2021

Computer Vision and Pattern Analysis Across Domains

November 24–25, 2021

University of Applied Sciences St. Pölten

Austrian Association of Pattern Recognition (OAGM)

Editors

Markus Seidl, Matthias Zeppelzauer, and Peter M. Roth

Layout

Austrian Association of Pattern Recognition
https://aapr.at/

Cover

Verlag der Technischen Universität Graz

Supported by

ISBN 978-3-85125-869-1
DOI 10.3217/978-3-85125-869-1

Contents

Preface

When we discussed the date for the 2021 edition of the OAGM workshop, we already had the rescheduling and finally cancellation of the 2020 edition in our minds. At the time of our discussion the shortage of COVID-19 vaccines was the dominant topic. Hence, when the decision between September and November was due, we decided that the workshop should take place on November 24 and 25, 2021 at University of Applied Sciences St. Pölten. Looking back, September would have allowed to hold the workshop as intended, namely To bring together researchers, students, professionals, and practitioners from the fields of Computer Vision and Pattern Recognition to present and actively discuss the latest research and developments.

Given the fast development of the pandemic situation, we had to cancel the workshop at the beginning of November. At this time, the review process was already finished. Consequently, it is possible to publish the conference proceedings. We thank the authors and reviewers for their contributions to this publication.

We received 21 original contributions of which 16 (10/12 full papers, 3/3 student papers, 3/6 spotlight papers) have been accepted, resulting in an overall acceptance rate of 76%. Each contribution was peer-reviewed in a double-blind process by at least two reviewers from an international program committee. One outstanding contribution will be awarded a best paper prize sponsored by OCG. In addition, there will be an IEEE Women in Engineering Award for the best contribution of a female first author. We want to thank OCG and IEEE for sponsoring these awards and Land Niederösterreich for the financial support of the unfortunately cancelled workshop and this publication. At the time of its cancellation, the workshop program was already planned. We thank the invited speakers Andreas Maier (FAU Erlangen-Nürnberg), Marc Pollefeys (ETH Zürich) and Markus Schedl (JKU Linz) for their willingness to give a presentation.

We hope to see you in 2022 at the next edition of the OAGM workshop, stay healthy,

Markus Seidl and Matthias Zeppelzauer (conference chairs)
Peter M. Roth (publication chair)
St. Pölten, December 2022

Workshop Chairs

Markus Seidl (FH St. Pölten)

Matthias Zeppelzauer (FH St. Pölten)

Publication Chair

Peter M. Roth (Technical University of Munich)

Program Committee

Alexander Schindler (Austrian Institute of Technology)
Andreas Uhl (University of Salzburg)
Arjan Kuijper (TU Darmstadt)
Armin Kirchknopf (St. Pölten University of Applied Sciences)
Christoph H. Lampert (IST Austria)
Csaba Beleznai (Austrian Institute of Technology, Austria)
Djordje Slijepcevic (St. Pölten University of Applied Sciences)
Florian Kleber (TU Wien)
Friedrich Fraundorfer (Graz University of Technology)
Gerhard Paar (Joanneum Research)
Gernot Stübl (PROFACTOR GmbH)
Gwenael Mercier (University of Vienna)
Harald Ganster (Joanneum Research)
Horst Bischof (Graz University of Technology)
Isabel Dregely (FH Technikum Wien)
Johannes Fürnkranz (JKU Linz)
Josef Scharinger (JKU Linz)
Levente Hajder (Eötvös Loránd University)
Margrit Gelautz (TU Wien)
Martin Kampel (TU Wien)
Martin Welk (UMIT Hall/Tyrol)
Mathias Lux (University of Klagenfurt)
Michael Bleyer (Microsoft)
Peter M. Roth (Graz University of Technology)
Robert Sablatnig (TU Wien)
Roland Perko (Joanneum Research)
Thomas Pock (Graz University of Technology)
Vincent Lepetit (ENPC ParisTech)
Walter G. Kropatsch (TU Wien)

OAGM Awards 2020

The

OAGM Best Paper Prize 2020

were awarded to the papers

Highly Accurate Binary Image Segmentation for Cars

by

Thomas Heitzinger and Martin Kampel.

and

Real-World Video Restoration using Noise2Noise

by

Martin Zach and Erich Kobler.

Index of authors

Evaluation of Monocular and Stereo Depth Data for Geometry-Assisted Learning of 3D Pose

Andreas Kriegler [1,2], Csaba Beleznai [1] and Margrit Gelautz [2]

***Abstract*— The estimation of depth cues from a single image has recently emerged as an appealing alternative to depth estimation from stereo image pairs. The easy availability of these dense depth cues naturally triggers research questions, how depth images can be used to infer geometric object and view attributes. Furthermore, the question arises how the quality of the estimated depth data compares between different sensing modalities, especially given the fact that monocular methods rely on a learned correlation between local appearance and depth, without the notion of a metric scale. Further motivated by the ease of synthetic data generation, we propose depth computation on synthetic images as a training step for 3D pose estimation of rigid objects, applying models on real images and thus also demonstrating a reduced synth-to-real gap. To characterize depth data qualities, we present a comparative evaluation involving two monocular and one stereo depth estimation schemes. We furthermore propose a novel and simple two-step depth-ground-truth generation workflow for a quantitative comparison. The presented data generation, evaluation and exemplary pose estimation pipeline are generic and applicable to more complex geometries.**

I. INTRODUCTION

Recent scientific trends increasingly allow for an enhanced spatial perception of a given environment and its actors. On one hand, this is partly facilitated by the recent surge in representational capacity and flexibility of learned representations. On the other hand, the emergence of enhanced depth-sensing modalities such as high-quality stereo vision, monocular depth estimation, LiDAR, Radar offer new geometry-encoding cues, which are highly invariant with respect to view, appearance and photometric variations. These spatial cues, along with appearance attributes, are often exploited in robotic perception and interaction tasks, such as pose-aware grasping and path planning.

3D object pose denotes the spatial transform needed to align the coordinate reference of an observed object with that of the observer. As depth data contains distinctive cues linked to the sought translational and rotational object pose parameters, in this paper we present a focused study on examining the data quality of monocular and stereo depth modalities in light of a learned pose estimation task.

A primary motivation of our work stems from the fact that models trained on synthetic data often exhibit a severe degradation when facing the real data domain [8], or learning requires a photorealistic pipeline [2] to close the gap between simulated and real data. To mitigate this problem, we propose depth computation from synthetic images, with the objective to derive a representation exhibiting less synthetic qualities. Depth data estimated from synthetic images (via monocular or stereo estimation schemes), however, still might convey specific characteristics, which limit generalization towards real-world situations. Therefore, we propose a depth-data-specific comparison based on the computed surface normals to examine how quality discrepancies of different depth modalities relate to each other. Furthermore, we also examine the use of such training data with synthetic origins in the context of learning 3D pose-aware detectors, as it is described later on.

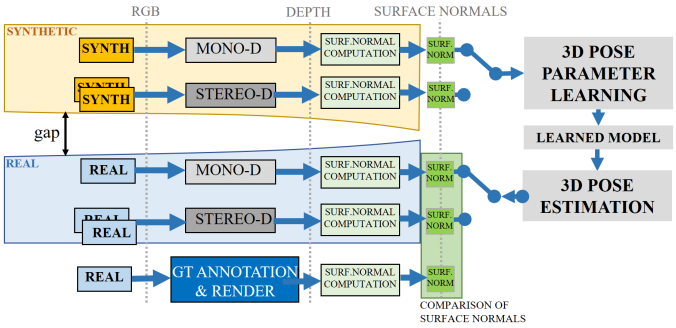


Fig. 1. Overview of the various depth and surface normals generation pipelines from synthetic and real data. Depth (computed surface normals) from synthetic images is used for training a pose-aware detector, which is tested on real images. Our proposed ground truth depth generation scheme is used to generate reference depth/surface normals data.

To support a quantitative comparison between the different depth modalities, we also propose a novel quantitative evaluation pipeline based on a simple ground truth generating procedure, yielding dense metric depth ground truth. Comparison of monocular depth estimates to a metric depth ground truth, however, is not straightforward due to the lack of metric scaling. To this end we propose an object-centered evaluation scheme, which compares computed surface normals at an object level and in a pixel-wise manner. Finally, to validate that a given transition to depth data narrows the synth-to-real gap, we present pose estimation experiments purely trained on depth from synthetic imagery. These experiments employ a baseline encoder-decoder-type pose estimation methodology and cylindrical objects as training and test objects. The presented data generation, evaluation and pose estimation scheme, however, is generic and also

*This work was partially carried out in the HOPPER project, supported by the "ICT of the Future" programme of the Austrian Research Promotion Agency (FFG)

[1]Assistive and Autonomous Systems, Center for Vision Automation and Control, AIT Austrian Institute of Technology, 1210 Vienna, Austria {andreas.kriegler, csaba.beleznai}@ait.ac.at

[2]Visual Computing and Human-Centered Technology, TU Wien Informatics, 1040 Vienna, Austria margrit.gelautz@tuwien.ac.at

applicable to more complex object geometries.

In summary, the paper proposes the following three contributions:

- Ground truth generation: we introduce a novel and generic annotation pipeline for computing dense and accurate depth and pose data for a wide variety of real scenes,
- Depth quality assessment: we propose a quantitative assessment scheme, comparing monocular and stereo-based estimates to ground truth via an object-centered analysis of computed surface normals,
- Initial results for 3D pose estimation trained on synthetic data: we demonstrate the feasibility of inferring 3D pose in real images via learned models trained on monocular and stereo depth normals, estimated from synthetic data.

The remainder of the paper is structured as follows: section II gives an overview on related work. Section III describes two data generation tasks: synthetic data generation for learning and depth ground truth generation for evaluation, both via Blender [3]. Section IV presents the proposed depth quality evaluation scheme. Finally, section V shows the applicability of a synthetic-data-based training pipeline to learn and predict object poses in real and synthetic images.

II. RELATED WORK

Recent research activities targeting learned representations of geometric traits encompass a large set of works, given that geometric shape and structure are intrinsic object properties which are highly invariant for different viewing and illumination conditions. This emerging field of geometric deep learning is well summarized in [4], [5], where geometric principles are highlighted to explain regularities often observed in the physical world, i.e. gravitational or right-angle structuring of man-made objects. Depth data naturally conveys geometric information, therefore understanding depth computation, its data characteristics and its failure modes are highly pertinent. [27] outlined four steps commonly encountered in classical stereo image pipelines. Despite representational advances via Deep Learning, these steps continue to play a key role [37]. Depth estimation from a single image, also denoted as monocular depth estimation, has recently emerged as an appealing alternative to depth estimation from stereo image pairs [36]. One of the first methods was [11] who also introduced scale-invariant evaluation metrics to measure the quality of the estimated depth maps. The proposed evaluation technique seeks an optimum depth scaling best aligning estimated depth and ground truth, a search step which is sensitive in presence of large depth discrepancies. Later works have explored continuously improving representations to learn a robust correlation between the appearance of a scene and its geometry [13], [22], [21]. Some works [20] approached this learning task as part of a multi-task learning scenario, where estimating the apparent motion and scene depth (from the viewpoint of a mobile observer) are formulated as two correlated and mutually-supporting learning tasks. An enhanced generalization of monocular depth models is attained via a mixture of datasets in [24]. Recent representational advances based on vision transformers [9], exploiting the attention-mechanism [32] are capable to accurately capture long-range semantic relations [23], see also [19] for a survey.

Nevertheless, the task of inferring absolute depth from a single image is an ill-posed problem, most prominently because of the prevailing scale ambiguity, making it unreliable in certain situations. These shortcomings lead to the conclusion in [28] that stereo vision is still required for accurate depth estimation, as stereo methods employ a principled, well understood, multi-view processing framework using concepts of the pinhole camera model [16]. In stereo vision, ambiguities are generated from other sources. In particular, stereo matching - that is, finding corresponding points in two (or more) stereo images for disparity estimation - is typically challenged by homogeneous image regions, repetitive patterns, depth discontinuities and occlusions, and particular surface reflectance properties. CNNs are known to be very powerful feature extractors and multiple learning-based deep neural network architectures exploit this capability for enhanced feature matching of stereo images, such as in [6], [12], [7]. AANet [34] provides a very good speed-accuracy trade-off.

While both monocular and stereo-based depth reconstruction methods have their individual advantages and limitations, one of the goals of our work is to provide a quantitative comparison of the two approaches in the context of a geometric deep learning task. Research works having a similar data characterization scope are still lacking. Although [28] provide a comparison between depths generated from monocular frameworks vs. stereo-setups, it is largely qualitative, and it does not use state of the art methods from the respective fields. [24] proposes several dataset-specific metrics, which are nevertheless difficult to relate across different datasets.

Finally, our paper is closely related to 3D object pose estimation from appearance and/or depth cues. Representational advances in recent years have resulted in an increasing accuracy and robustness with respect to clutter, occlusions and pose-ambiguous object types. This evolution is prominently reflected in the BOP Challenge series [18]. This paper leans on its resulting insights, that data availability and the domain gap between synthetic training and real test often represent a hurdle. These findings motivate us to devise data generation schemes which yield data conveying spatial cues and better bridge the domain gap.

III. GROUND TRUTH FOR DEPTH AND POSE

In the present data-driven era of Deep Learning, model performance is closely linked to the quantity and quality of training data [24]. The generation of synthetic data using GANs [14] has proven successful, while the inclusion of data from different sources such as YouTube videos [1] or movie datasets [24] is gaining interest as well. Nevertheless, the exploitation of frameworks commonly used in computer graphics applications is still largely unexplored [25]. One such program is Blender [3], which is commonly used to

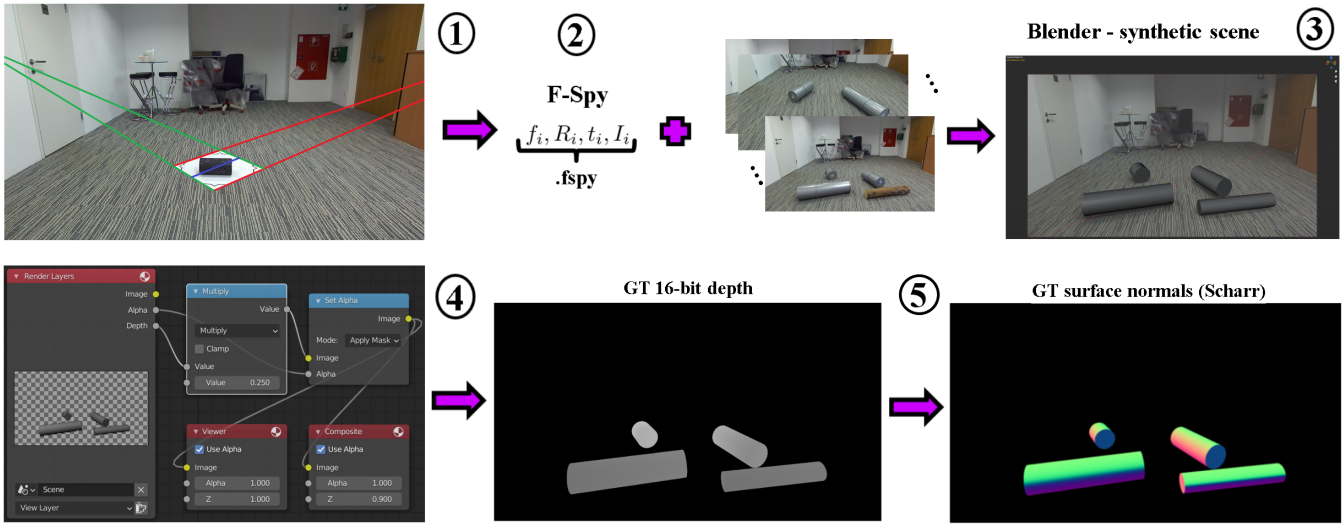


Fig. 2. Our proposed 3D object annotation pipeline yielding synthetically-correct per-pixel depth values and object 6DoF poses for real images captured with an RGB camera. The pipeline comprises five steps: 1) Alignment of two sets of parallel lines (red and green) where the sets are orthogonal, forming the ground plane. Additionally, a line segment of known length (blue) is set. 2) Creation of synthetic camera with estimated intrinsic and extrinsic parameters. 3) Camera and frames to be annotated are imported in Blender to create a synthetic twin of the scene. 4) Rendering of 16bit depth images. 5) Final transform of depth maps to surface normal images.

render synthetic imagery based on modelled or procedurally-generated scenes. In this work the use of the Blender platform is two-fold: training data generation for learning pose-aware object detectors, and 3D scene annotation for generating dense depth ground truth. These functionalities are explained in detail below:

Synthetic data generation: We procedurally generate a large synthetic dataset (consisting of cylindrical objects) along with 3D pose annotations. Diversity is introduced in form of various spatial object configurations and varying view parameters. We generate a rich set of object-specific annotations in form of 6DoF pose parameters, metric object dimensions, 2D bounding box, center point location and occlusion indicators using ray-tracing. Example renders can be seen at the top left side of Fig. 3, with textures randomized for each view. Please note that texture plays only a minor role as the RGB domain is not used directly for learning. Random textures, on one hand, generate a notion of the governing perspective and corresponding locations, thus facilitating the task of monocular and stereo depth estimation. Furthermore, random textures also introduce small-scale texture-induced monocular depth artifacts, thus robustifying a learned model with respect to locally corrupted depth/normal data. 56k of such synthetic samples (resolution $768\times512px$) are used to train and validate our object location and pose regression models, as described in Section V.

3D scene calibration, ground-truth-depth generation: Trained pose-aware models shall be evaluated on a real image dataset. To this end, we capture 120 rectified stereo image pairs using a Stereolabs ZED2 stereo camera [30] with a resolution of $1920\times1080px$. Captured images depict four different office environments (further on denoted as *scenarios*) with a variable number of cylindrical objects lying on a common ground plane. To generate a dense reconstruction for all scenes with as little effort as possible, we rely on a simple photogrammetric concept. We employ the technique by Guillou et al. [15], requiring two vanishing points and a line segment of known length. The two vanishing points can be easily defined by two line pairs, pairwise orthogonal to each other in the real world. Given these inputs, the camera rotational and translational parameters can be determined, along with its focal length. To perform the calibration and scene reconstruction, we execute following steps:

- Calibration: for a given scenario, we assume a stationary camera mounted on a tripod. We create a blank rectangular (cardboard) shape with at least one known dimension to use as a calibration target (see Step 1 in Fig.2). In the very first frame of a given scenario, the edges of the calibration target can be used to manually delineate two pairs of parallel line segments, yielding the sought camera parameters. In later images of a given scenario the target can be removed, since the camera views remain stationary. The publicly available fSpy toolkit [31] offers an interactive interface for the calibration algorithm [15]. It also provides a Blender camera generator functionality to create an equivalent camera within Blender, adequately oriented, translated and scaled, existing within a 3D space defined by the line segments leading to the vanishing points.
- 3D scene annotation in Blender: in this step we would like to spatially align a number of geometric objects within the 3D metric space of the camera. Each image of a scenario contains N ($1<N<4$) cylindrical objects, randomly placed in a lying pose. We measure the dimensions (length, radius) of these objects, in order to create cylindrical primitives of the same metric size in Blender. If objects are known to be on a common ground plane, object dimensions are not necessary.

Nevertheless, known dimensions significantly constrain the possible pose space where placed 3D objects are aligned with the apparent projections in the view seen through the camera (Steps 2 and 3 in Fig.2). This step also naturally leads to object pose attributes (orientation, translation) with respect to the camera.

- Dense ground-truth-depth generation: after aligning all 3D objects within Blender, the scaled Z-depth information can be rendered for the given scene. This depth information contains the metric depth for every scene point, similarly to a depth measurement via calibrated stereo cameras. This step is executed programmatically and produces float-valued depth entries for every image point where camera rays hit a previously placed object (see Step 4 in Fig.2). Note that a scaling factor is introduced for visualization only.

The presented calibration and ground truth generation scheme represents a straightforward annotation workflow for 3D object pose and depth data. It is applicable for a wide range of object geometries, scale ranges and arbitrary viewpoints. A detailed documentation, sample scene and code can be found at [*url*].

IV. EVALUATION OF MONOCULAR AND STEREO DEPTH DATA

In this section we describe a data-oriented evaluation methodology for three state-of-the-art depth estimation schemes. Our comparison targets the quality evaluation of structure-encoding surface normals (derived from the depth data), in the context of representation learning for 3D pose estimation. The three selected estimation schemes consist of two monocular depth estimation methods MiDaS_v2.1 [24] and MiDaS_v3.0 [23], and a stereo depth estimation model AANet [34]. Further on, MiDaS_v2.1, MiDaS_v3.0 (using the *DPT_large* model) and AANet are denoted as MiDaS, DPT and AANet for brevity.

MiDaS [24] is a CNN-based depth estimator using the framework of [33] with a ResNet [17] backbone. It was trained using up to ten different datasets, including 3D movies, leveraging the large data quantity and diversity for generalization. DPT [23] is a vision transformer trained for multiple dense prediction tasks including depth estimation. An argument for transformers is that they are able of capturing long-range semantic relationships in images [9], which should enforce stronger global structural consistency in the depth results; a trait which is often lacking for CNN-based monocular-depth frameworks [23]. Lastly, AANet [34] consists of an adaptive aggregation model for multi-scale disparity cost aggregation, resulting in an efficient stereo matching scheme. We employ the AANet *kitti*2015+ model which incorporates GANet [35] for feature matching.

A direct pixel-wise comparison between a monocular depth estimate and a ground truth is not straightforward, as monocular-depth frameworks generate disparity (inverse depth) values with no metric scaling. Furthermore, different monocular models yield disparity (depth) estimates with substantially different scaling factors. Therefore, common stereo vision evaluation metrics - assuming data living in a metric space - cannot be applied. To overcome this problem, we propose an object-centered scaling scheme performing a normalization within object-specific regions. The objective of this scaling step is to bring monocular, stereo and ground truth depth data within object foreground regions into a scale-normalized form. The input for this scaling is a depth image D_i, where $i = \{0,1,2\}$ indicates ground truth, monocular and stereo depth, respectively. A corresponding object foreground mask m_0 is also needed (generated via the ground truth generation process) to spatially constrain the set of pixels included into the normalization step. This mask contains unit entries at all object locations, denoted as object mask region m_0^{Obj}, and zeros elsewhere. The scaling operation is performed as:

$$D_i^* = D_i \,/\, max(1, D_i[m_0^{Obj}]), \tag{1}$$

resulting in D_i^*, a scaled depth image containing unit-normalized values within the object mask region. This subset of normalized depth values is used exclusively for further computation steps towards a quantitative comparison.

We adopt this object-foreground-based normalization procedure for ground truth, stereo and monocular depth data, resulting in depth values within the object foreground regions scaled to a common range. Instead of using the scaled depth values for an evaluation, we investigate its spatial derivatives, in form of surface normals. The choice of opting for surface normals stems from a representational consideration: when seeking to learn representations for objects situated at varying distances from a camera, computed surface normals exhibit less variation than depth data. We use a simple procedure to transform depth images to surface normals. First we calculate pixel intensity changes as derivatives d_x and d_y using the Sobel kernel [29]. With the intensity gradients we build local support planes, whose normal vectors can be seen as the normal vectors of the object surface in those pixels. As matrix norm we use the Frobenius norm. Since the Sobel kernel size is a configurable parameter, most commonly 3×3, 5×5 or 7×7, we examine resulting surface normal quality variations in this regard, and also include an alternative Scharr 3x3 kernel [26]. Images of computed surface normals are visualized by mapping the directional vector components to respective 8-bit RGB channels. The 3D vectors of surface normals are compared to ground truth in a pixel-wise manner using a 3D cosine similarity, yielding a similarity score in the range of $[-1,1]$. Vector similarity scores are mapped to a $[0,1]$ range and a cumulative score from all object foreground regions is formed.

V. RESULTS AND DISCUSSION

In this section, we present results on comparing depth data quality in terms of pixel-wise surface normal similarities with respect to a corresponding ground truth. The comparison is generated for a real dataset using the presented three (MiDaS, DPT and AANet) depth computation modalities. In addition to a quantitative evaluation, we also present qualitative results on the depth quality and experimental outcomes for 3D pose

estimation. In the following, we describe our dataset and related evaluation results, followed by qualitative depth and pose estimation results.

Dataset and evaluation results: our 120 real-image dataset (see Section III) was captured using 4 distinct viewpoint setups, each scene containing 30 random object configurations. For each of the 120 images a corresponding depth ground truth was computed. Table I displays surface normal similarities computed from the MiDaS, DPT and AANet methods with respect to the ground truth. The table also shows the effect of the varying kernel-size used for surface normal computation. The kernel size of -1 denotes the Scharr, the other numbers relate to the Sobel kernel size. As it can be seen from the table, the two monocular depth estimation methods on an average produce very similar quality. While DPT tends to produce more geometric detail, in case of our targeted, smoothly varying surfaces it did not lead to enhanced scores. On the other hand, the AANet stereo matching scheme clearly outperforms the monocular models in all cases. The elongated cylindrical objects are long enough to call for the need of estimating accurate far-range structural correlations; a trait where monocular methods are still lacking behind the quality of the stereo depth data. Monocular methods generate a spatially-smooth output, where derivative kernels of increasing size deteriorate the captured geometric details. Therefore, a small kernel size of $3px$ seems to produce optimum results. AANet, on the other hand, benefits from larger kernels, suppressing the noise associated with the disparity estimation process.

Qualitative depth results: Fig. 3 displays a large set of qualitative results, partially generated for synthetic renders (Fig. 3 left half), partially for our real-image dataset. To facilitate the interpretation of depth quality, besides false-color depth images we also display two views of the point cloud representing the given scene. As it can be seen from the point cloud views, monocular depth estimation is relatively accurate when considering it locally, but at a large scale (especially near the image boundaries) significant spatial deviations occur. This observation also implies that, if pursuing an object detection or pose estimation task, local depth cues from monocular estimation can provide valuable hints, lifting many ambiguities associated with the monocular nature of the view. However, for problems requiring a global scale consistency, stereo pipelines still seem to be the more accurate and less data-dependent choice.

Qualitative pose estimation results: To examine the influence of data quality onto a 3D pose estimation learning task, we performed following experiments. Section III describes our data generation step for learning. The synthetic data, as single images or stereo pairs, are used within the respective monocular (MiDaS, DPT) and stereo (AANet) pipelines to generate depth data. The computed surface normals from depth and corresponding pose annotations $\{$*class, 2D center, depth, angular parameters*$\}$ represent the input of our learning scheme. Given these inputs, an encoder-decoder-type framework (CenterNet [10]) was used to train depth-modality-specific models for 3D pose estimation (see also Fig. 1). We observe fast convergence within 3-5 epochs.

TABLE I

QUANTITATIVE COMPARISON OF SURFACE NORMALS COMPUTED USING DEPTH CUES FROM THREE MODELS. (HIGHER IS BETTER).

	k	*MiDAS_v2.1*[24]	*DPT_Large*[23]	*AANet*[34]
Scene 1	-1	0.7861	0.7898	0.8801
	3	0.7861	0.7898	0.8801
	5	0.7846	0.7887	0.8888
	7	0.7826	0.7870	0.8922
	avg.	0.7849	0.7888	**0.8853**
Scene 2	-1	0.8596	0.8722	0.9291
	3	0.8596	0.8722	0.9291
	5	0.8579	0.8714	0.9362
	7	0.8559	0.8702	0.9387
	avg.	0.8583	0.8715	**0.9333**
Scene 3	-1	0.8382	0.8005	0.8930
	3	0.8382	0.8005	0.8930
	5	0.8364	0.7984	0.9012
	7	0.8341	0.7958	0.9041
	avg.	0.8367	0.7988	**0.8978**
Scene 4	-1	0.8650	0.8604	0.9212
	3	0.8650	0.8604	0.9212
	5	0.8630	0.8588	0.9294
	7	0.8606	0.8567	0.9324
	avg.	0.8634	0.8591	**0.9261**

We demonstrate the applicability of our process to learn 6DoF pose parameters from purely synthetic data and perform prediction in real images. Fig. 3 shows pose estimation results (rows 6, 11 and 16) for the individual depth modalities. As it can be seen from the figure, in the synthetic domain results exhibit a high recall and a high pose estimation accuracy. In the real domain, however, inference on surface normals from monocular techniques shows several failure modes, a lower recall and precision, in form of occasionally hallucinating cylindrical objects within the nearby structural clutter. When using stereo-depth-based surface normals, however, results improve. Recall is still lacking, but no objects are hallucinated. Based on these results we believe that spatial cues derived from synthetic images can represent a way towards learning geometry-aware representations of objects and pose, which also exhibit validity within the real world.

VI. CONCLUSIONS

We present a geometrically-inspired depth data analysis scheme comparing surface normal cues from monocular and stereo-based pipelines, with object detection and 3D pose estimation tasks in mind. To support the data evaluation task, we propose a novel ground truth generation scheme, where dense depth and pose data can be created with little manual interaction. Our evaluations with respect to ground truth indicate that stereo-depth prevails in terms of data quality when compared to monocular depth, especially if a long-range depth data consistency is required. However, we demonstrate, that monocular depth still captures relevant local geometric details, which is sufficient to learn pose-aware object detectors from purely synthetic data. The demonstrated transition from the synthetic to the real domain seems to offer further geometry-aware analysis perspectives, while exploiting monocular or stereo depth cues.

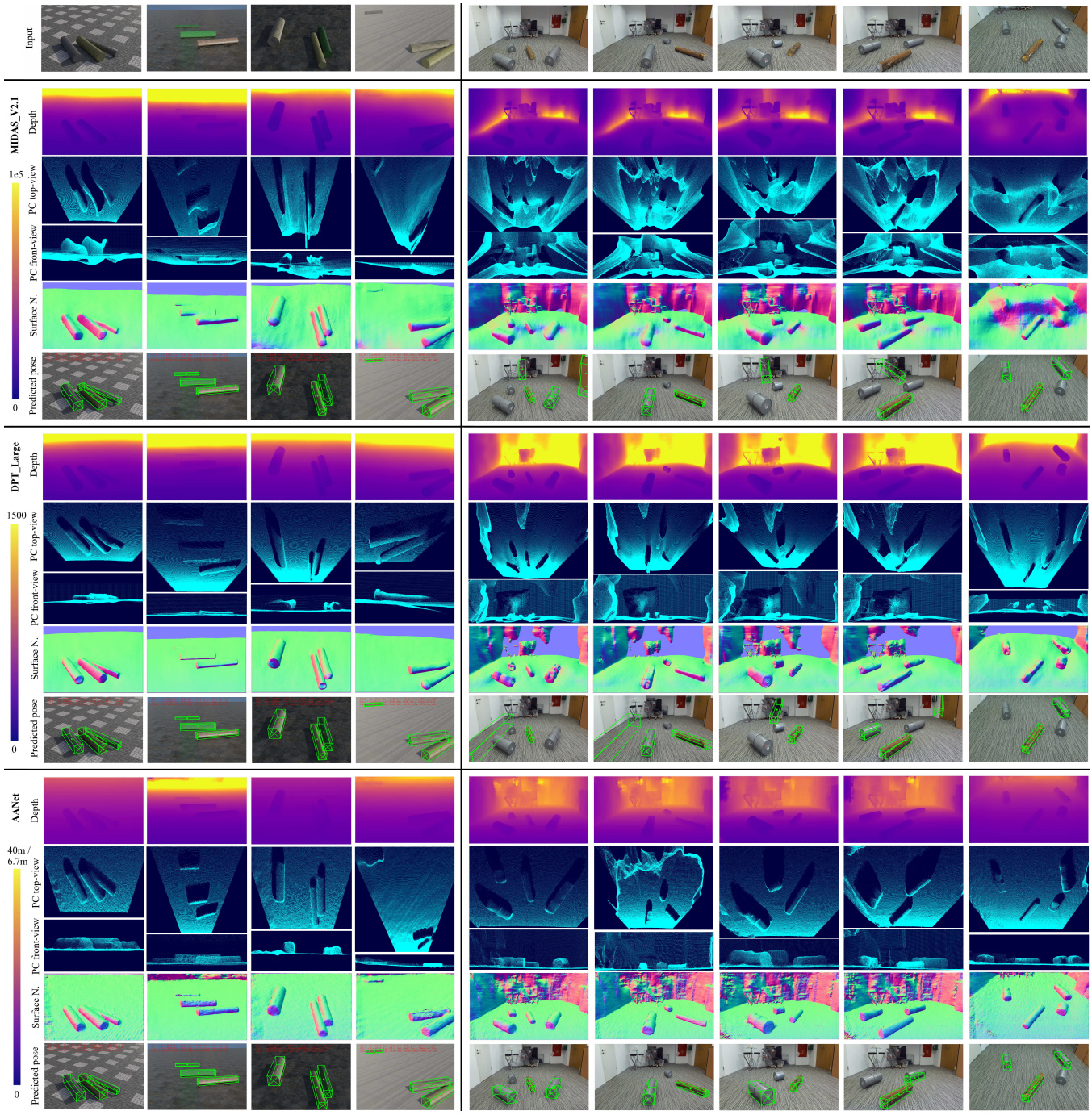


Fig. 3. For the synthetic data domain (left) and real images (right) we visualize example input images as well as depth images, depth point clouds and surface normals obtained using each of three depth estimation methods. The final rows for each model show CenterNet pose estimation results trained on surface normals from the large-scale synthetic dataset.

REFERENCES

[1] S. Abu-El-Haija, N. Kothari, J. Lee, A. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *ArXiv*, vol. abs/1609.08675, 2016.

[2] H. Alhaija, S. Mustikovela, L. Mescheder, A. Geiger, and C. Rother, "Augmented reality meets computer vision: Efficient data generation for urban driving scenes," *International Journal of Computer Vision (IJCV), 2018*, 2018.

[3] *Blender - a 3D modelling and rendering package*, Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. [Online]. Available: http://www.blender.org

[4] M. Bronstein, J. Bruna, T. Cohen, and P. Velivckovic, "Geometric deep learning: Grids, groups, graphs, geodesics, and gauges," *ArXiv*, vol. abs/2104.13478, 2021.

[5] W. Cao, Z. Yan, Z. He, and Z. He, "A Comprehensive Survey on Geometric Deep Learning," *IEEE Access*, vol. 8, pp. 35 929–35 949, 2020.

[6] Z. Chen, X. Sun, L. Wang, Y. Yu, and C. Huang, "A Deep Visual Correspondence Embedding Model for Stereo Matching Costs," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 972–980.

[7] X. Cheng, Y. Zhong, M. Harandi, Y. Dai, X. Chang, H. Li, T. Drum-

mond, and Z. Ge, "Hierarchical neural architecture search for deep stereo matching," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 22 158–22 169.

[8] C. Doersch and A. Zisserman, "Sim2real transfer learning for 3d human pose estimation: motion to the rescue," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ArXiv:2010.11929*, p. 21, 2020.

[10] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint Triplets for Object Detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 6569–6578.

[11] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in Neural Information Processing Systems*, vol. 3, 2014, pp. 2366–2374.

[12] J. Flynn, I. Neulander, J. Philbin, and N. Snavely, "DeepStereo: Learning to Predict New Views from the World's Imagery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 842–857.

[13] R. Garg, B. G. Vijay Kumar, G. Carneiro, and I. Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9912 LNCS, no. April, pp. 740–756, 2016.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, p. 139–144, Oct. 2020.

[15] E. Guillou, D. Méneveaux, E. Maisel, and K. Bouatouch, "Using vanishing points for camera calibration and coarse 3d reconstruction from a single image," *Vis. Comput.*, vol. 16, no. 7, pp. 396–410, 2000.

[16] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, ISBN: 0521540518, 2004.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[18] T. Hodaň, M. Sundermeyer, B. Drost, Y. Labbé, E. Brachmann, F. Michel, C. Rother, and J. Matas, "BOP challenge 2020 on 6D object localization," *European Conference on Computer Vision Workshops (ECCVW)*, 2020.

[19] S. Khan, M. Naseer, M. Hayat, S. Waqas Zamir, F. Shahbaz Khan, and M. Shah, "Transformers in vision: A survey," *ArXiv:2101.01169*, p. 28, 2021.

[20] C. Luo, Z. Yang, P. Wang, Y. Wang, W. Xu, R. Nevatia, and A. Yuille, "Every Pixel Counts ++: Joint Learning of Geometry and Motion with 3D Holistic Understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2624–2641, 2020.

[21] S. Mahendran, "Geometric Deep Learning for Monocular Object Orientation Estimation," Ph.D. dissertation, Hopkins University, 2018.

[22] R. Mahjourian, M. Wicke, and A. Angelova, "Unsupervised Learning of Depth and Ego-Motion from Monocular Video Using 3D Geometric Constraints," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5667–5675, 2018.

[23] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," *arXiv:2103.13413*, p. 15, 2021.

[24] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 14, 2020.

[25] S. Reitmann, L. Neumann, and B. Jung, "BLAINDER-A Blender AI Add-On for Generation of Semantically Labeled Depth-Sensing Data," *Sensors*, vol. 21, no. 6, 2021.

[26] H. Scharr, "Optimal filters for extended optical flow," in *IWCM*, 2004.

[27] D. Scharstein and R. Szeliski, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms," Microsoft Research, Tech. Rep., 2001.

[28] N. Smolyanskiy, A. Kamenev, and S. Birchfield, "On the Importance of Stereo for Accurate Depth Estimation: An Efficient Semi-Supervised Deep Neural Network Approach," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 11 200–11 208, 2018.

[29] I. Sobel and G. Feldman, "Sobel - isotropic 3 x 3 image gradient operator," A Talk at the Stanford Artificial Project, 271-272, 1968.

[30] StereoLabs. (2019) Zedcam2. Stereo Labs. [Online]. Available: https://www.stereolabs.com/zed-2/

[31] Stuffmatic, "fSpy." [Online]. Available: https://fspy.io

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017, pp. 6000–6010.

[33] K. Xian, C. Shen, Z. Cao, H. Lu, Y. Xiao, R. Li, and Z. Luo, "Monocular relative depth perception with web stereo data supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[34] H. Xu and J. Zhang, "AANet: Adaptive Aggregation Network for Efficient Stereo Matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[35] F. Zhang, V. Prisacariu, R. Yang, and P. H. S. Torr, "Ga-net: Guided aggregation net for end-to-end stereo matching," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 185–194, 2019.

[36] C. Zhao, Q. Sun, C. Zhang, Y. Tang, and F. Qian, "Monocular depth estimation based on deep learning: An overview," *Science China Technological Sciences*, pp. 1–16, 2020.

[37] K. Zhou, X. Meng, and B. Cheng, "Review of Stereo Matching Algorithms Based on Deep Learning," *Computational Intelligence and Neuroscience*, vol. 2020, no. 1, 2020.

An Evaluation of the Machine Readability of Traffic Sign Pictograms using Synthetic Data Sets*

Alexander Maletzky[1], Stefan Thumfart[1] and Christoph Wruß[2]

***Abstract*— We compare the machine readability of pictograms found on Austrian and German traffic signs. To that end, we train classification models on synthetic data sets and evaluate their classification accuracy in a controlled setting. In particular, we focus on differences between currently deployed pictograms in the two countries, and a set of new pictograms designed to increase human readability. We find that machine-learning models generalize poorly to data sets with pictogram designs they have not been trained on, and conclude that manufacturers of advanced driver-assistance systems (ADAS) must take special care to properly address small visual differences between different traffic sign pictogram designs. Our main contributions are the creation of a vast synthetic data set of traffic sign images, training and evaluating state-of-the-art classification models to assess the machine readability of different pictogram designs, and employing techniques from explainable AI to analyze which image regions are particularly important to the classifiers.**

I. INTRODUCTION

In recent years, the number of semi-autonomous vehicles and advanced driver-assistance systems (ADAS) on our streets has been growing steadily. Even if there are still a lot of problems to be resolved before machines can eventually take over entirely, certain aspects of driving have been successfully automated already. One of them is *traffic sign recognition*, which consists of *detecting* and *classifying* traffic signs in video frames produced by a forward-facing camera. The results of this recognition process can then be used to automatically control the speed of the vehicle, or to display the found traffic signs on the instrument panel to inform the driver about them. In either case, correctly recognizing the traffic signs is of paramount importance for avoiding potentially fatal accidents. In the long-term future human-readable traffic signs will maybe disappear entirely, but in the current mixed-traffic regime machines must still be able to recognize traffic signs tailored to human needs.

State-of-the-art convolutional neural networks (CNNs) achieve near- or even super-human performance in many computer vision benchmark tasks, including traffic sign recognition [25]. However, as prior works illustrates, they may at the same time fail to correctly classify input that slightly deviates from the training distribution [36], [34], [15], [12], [7]. In our experiments we seek to find out whether and how this observation applies to *traffic sign classification models* under varying *pictogram designs*. Concretely, we pose the following questions: (i) Are there significant differences concerning the machine readability of different pictogram designs? In particular, we compare the current Austrian and German designs, as well as a proposed new Austrian design. (ii) How well do models generalize from one design to a new, unseen design? (iii) Which image details and regions are particularly important to classification models, and can this information be used to derive design rules that improve machine readability?

For answering these questions we trained traffic sign classifiers on a vast *synthetic data set*. The reason why we used synthetic- rather than real-world data is twofold: (i) A fair, systematic comparison of the machine readability of different pictogram designs is difficult to realize on real-world data with inherent differences *besides* the actual pictogram design. (ii) No real-world data exists for the proposed new Austrian design. The latter point is of particular importance, because whenever a traffic sign (pictogram) design is replaced by a new one, existing ADAS must be tested on and possibly adapted to the new design despite the lack of real-world training data. Our work demonstrates how this can be accomplished with carefully-crafted synthetic data sets.

An extended version of this paper can be found on arXiv [27].

A. Related Work

There exists a large body of scientific work regarding the automatic detection and classification of traffic signs in real-world as well as synthetic data sets. One of the most widely used real-world data sets is the German Traffic Sign Recognition Benchmark (GTSRB) [35] for classifying small image patches extracted from traffic scenes into one of 43 classes. Similar data sets exist for traffic signs from other countries and territories [29], [16], [43], [37], [40], [18], [8], [24], [42], [31], [14].

Closer to the kind of data sets employed in our experiments are partly synthetic data sets, where photographs or video frames of full traffic scenes are augmented with ultra-realistic weather effects [33], [10], [41]. In addition to these partly synthetic data sets there also exist data bases of fully synthetic 3D renderings of traffic scenes under varying (weather) conditions [4], [32], [2], [3], [5]. All these data sets have in common that they are better suited for object *detection* tasks, though. In [39], [28] and [9], real-world traffic scenes and signs are systematically modified by adding weather effects and other types of corruptions, to evaluate

*This research was funded by FFG (Austrian Research Promotion Agency) under grant 879320 (SafeSign) and supported by the strategic economic research programme "Innovatives OÖ 2020" of the province of Upper Austria.

[1]A. Maletzky and S. Thumfart are with RISC Software GmbH, 4232 Hagenberg, Austria. alexander.maletzky@risc-software.at

[2]Ch. Wruß is with ASFINAG Service GmbH, 1230 Vienna, Austria

Fig. 1: Overview of the experimental setup. Starting from three sets of traffic sign pictograms (each of a different design) a large collection of embedded and corrupted images (pictogram + traffic sign + background) is created. These images are then used to train classification models. Comparing the performance of the models allows to draw conclusions about the machine readability of the initial pictograms.

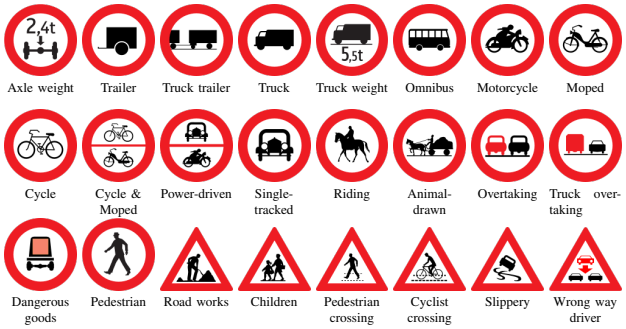


TABLE I: List of the 24 selected traffic sign classes, in the current Austrian pictogram design.

how well traffic sign detectors/classifiers work under such 'challenging conditions'. On the one hand, this resembles the approach we take in our experiments, but on the other hand, the main goal of the cited works is to compare different corruption types, not traffic signs or pictograms.

Similarly, in [20] a corrupted and perturbed version of ImageNet [11] is created. The methods employed there for corrupting images are similar to ours. ImageNet, however, is a general image data base without any particular focus on traffic signs, and the goal of [20] is to evaluate the performance of classification models in general, comparing models trained on 'clean' images to models trained on corrupted versions thereof.

II. METHODS

Fig. 1 presents an overview of the experimental setup. We first created synthetic data sets with traffic sign images and then trained classification models on them. Finally, we evaluated and compared the classification accuracy of these models.

A. Creating the Synthetic Data Sets

For creating the traffic sign images in the synthetic data sets, we started from high-resolution photographs of traffic scenes on Austrian highways in the year 2014 and extracted 14 patches with traffic signs. Seven of these 14 patches contain a prohibitory sign (round with red border), the other seven contain a warning sign (triangular with red border).[1] We then analyzed each of these 14 images w. r. t. color spectrum and perspective, obtaining parameters that allow to automatically replace the displayed pictogram by any given new pictogram in a way that makes the resulting image still look realistic. We then doubled the number of images by flipping them horizontally.

Next, we selected 24 traffic sign classes of the 'prohibitory' (18) and 'warning' (6) categories for our experiments. Pictograms of the current Austrian and German design could simply be downloaded from [6] and [1], respectively. Four of the 24 selected classes exist in Austria but do not have a German counterpart, meaning that we had to craft the corresponding pictograms manually by combining elements from other German pictograms. Pictograms of the proposed new Austrian design were kindly provided by their designer.[2] The complete list of classes is shown in Table I. Note that the selection of the 24 classes was mainly driven by the availability of a new Austrian pictogram design.

We replaced the pictograms in the 28 source images by the pictograms of the 24 selected classes, giving rise to a set of 336 images per pictogram design, with 14 images per class. We resized these images to a uniform size of 64×64 pixels. Finally, we augmented the set of 336 images by applying an arsenal of augmentation methods with varying intensities [21]. In particular, first one out of ten preselected corruption methods, like Gaussian noise, blurring, rain patterns, etc. is applied. Then, the resulting images are down-sampled by first down- and then up-scaling them, to decrease their spatial resolution but keep the size of 64×64 pixels; no extra smoothing is applied. The purpose of down-sampling is to simulate distance, as one of the key properties of well-designed traffic sign pictograms is being readable from large distances. We generated 250 variants for each of the 336 clean images, with five different levels of corruption intensity (50 per level). These intensities only affect the down-sampling factor, i. e., a higher intensity level gives rise to more 'pixelated' images.

Fig. 2 summarizes the whole data generation process. In the lower-left corner, two images per corruption intensity are shown, with intensity increasing from left to right. Eventually, every data set consists of 84,000 images, which are equally distributed across source patches (12,000 per patch), pictogram classes (3,500 per class) and corruption intensity (16,800 per intensity level). This, however, only corresponds to *one* data set, for one pictogram design. Repeating the process outlined above for each of the three designs yields three data sets with 252,000 images, where by construction identical corruptions are applied to the images of each design to enable an unbiased comparison. In order to obtain reliable results and reduce the impact of the randomness inherent to data augmentation on our results, we repeated the entire data generation process as well as the subsequent model training and evaluation three times and then averaged all results over these three independent runs. In total **756,000 images** were

[1] The source images and patches can be provided upon request.

[2] Stefan Egger, https://visys.pro/.

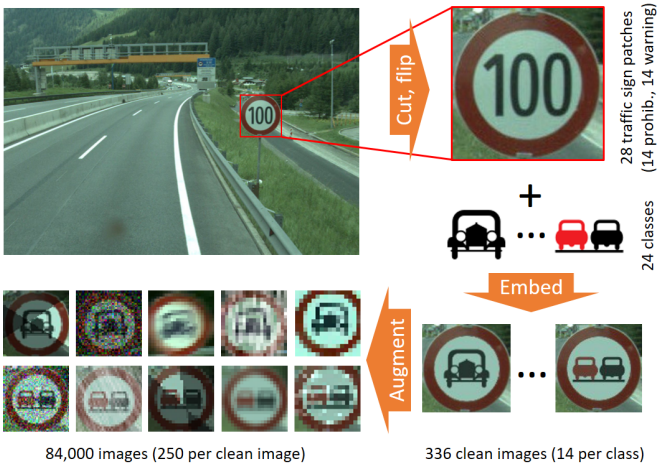


Fig. 2: Data generation process for the synthetic data sets used in our experiments. This process is repeated three times for current Austrian pictograms, proposed new Austrian pictograms, and current German pictograms, yielding nine data sets with a combined total of 756,000 images.

generated for our experiments.

For the sake of brevity, the data set with current Austrian pictogram design will be labeled AT_c, the one with the proposed new Austrian design will be labeled AT_n, and the one with the current German design will be labeled DE in the remainder. The combined data set with all currently deployed designs, i. e., the union of AT_c and DE, will be labeled CUR.

B. Model Training and Evaluation

The classification models were trained separately on each of the three pictogram designs (AT_c, AT_n, DE), as well as jointly on the two current designs (CUR). We considered two deep neural network architectures: a small ResNet architecture [19] with 20 layers and an input size of 64×64 pixels, and the architecture by Li and Wang [25] with an input size of 48×48 pixels. The latter was a natural choice for our experiments, since it represents the state-of-the-art on the GTSRB data set [35], with 99.66% test accuracy.

We split the data into training-, validation- and test sets and trained the models for 60 epochs, using the Adam optimizer [22] with an initial learning rate of 0.001, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is reduced by 80% whenever the validation loss does not improve for ten epochs. In the end, the trained weights of the epoch with the smallest validation loss are taken. Both training- and validation accuracy plateau after only a few (< 10) epochs in each case, so training for a total of 60 epochs is certainly sufficient. The splits into the three sets are based on the 28 source patches all images ultimately originate from, and are identical for each pictogram design.

After training, all models are evaluated on the held-out test sets, using the overall classification accuracy as the main metric of interest. As is common practice, confusion between classes is treated uniformly. Putting less weight on confusion between semantically similar classes (e. g., 'Pedestrian crossing' and 'Cyclist crossing') could be an interesting direction for future research.

	Li-Wang [25]	ResNet [19]
AT_c-AT_c	98.89±0.11	98.48±0.35
AT_n-AT_n	98.68±0.17	98.45±0.09
DE-DE	98.85±0.17	98.23±0.56
CUR-CUR	98.69±0.06	98.28±0.10
AT_c-AT_n	80.18±0.97	77.76±3.97
AT_c-DE	83.94±0.92	80.43±2.88
AT_n-DE	75.33±0.31	74.24±1.07
DE-AT_c	82.03±1.69	77.26±0.26
DE-AT_n	77.35±1.52	72.82±0.82
CUR-AT_n	85.48±1.27	84.17±0.60

TABLE II: Test accuracy (%) of the models trained in our experiments, displayed as *mean*±*SD* over three runs.

First, every model is evaluated on its 'own' test set, i. e., with the same pictogram design as in the set it was trained on. Due to the uniform construction of training-, validation- and test sets, the performance scores thus obtained are feasible for comparing the quality of different models, even if they are trained and evaluated on different pictogram designs. Besides evaluation on the own test set, some models are evaluated on 'foreign' test sets with different pictogram designs, too, to find out how well they generalize to unseen designs. In the remainder, evaluations will be denoted by short identifiers like AT_c-AT_n, where the data set label before the dash indicates the design the model was *trained* on, and the data set label after the dash indicates the design it was *evaluated* on.

III. RESULTS

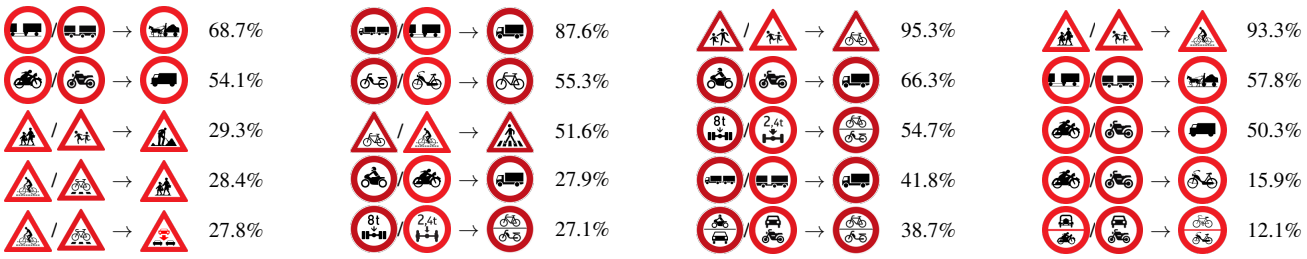
Table II shows the classification accuracy of all models. One can see that there is hardly any difference in the classification accuracy of the models between the three pictogram designs (top-three rows in Table II), and that the Li-Wang models generally tend to outperform the corresponding ResNet models by a small margin.

One can also see very clearly that the classification accuracy of every model drops significantly when evaluated on a 'foreign' test set, with different (albeit similar) pictograms. In fact, the difference between current and proposed new Austrian pictograms seems to be more pronounced than the difference between current Austrian and German pictograms. Models trained on German pictograms generalize only poorly to new Austrian pictograms, and vice versa; this is particularly interesting, since intuitively the design of the new Austrian pictograms resembles the German design much closer than the current Austrian design does, especially w. r. t. stroke width and level of detail.

A. Per-Class Results for Foreign Test Sets

We focus on the Li-Wang models [25] in the remainder of this section. The results of the ResNet models exhibit the same overall tendency as the Li-Wang models, including the frequently confused classes.

Table III lists the pairs of traffic sign classes the models confuse most often if the pictogram design differs from the



(a) AT_c-AT_n (b) DE-AT_c (c) DE-AT_n (d) CUR-AT_n

TABLE III: Frequent confusion of the models when evaluated on 'foreign' test sets. The numbers on the right are the percentages of samples belonging to the class on the left-hand-side of the arrows, which are misclassified as the class on the right-hand-side of the arrows. For better comparison, both training- and evaluation pictograms are shown on the left-hand-side of the arrows; in the case of CUR-AT_n, only current Austrian pictograms are shown, although German pictograms are part of the training set, too.

	AT_c-AT_n	DE-AT_c	DE-AT_n	CUR-AT_n
	11.9% (24)	1.1% (24)	15.7% (22)	26.9% (22)
	29.2% (22)	56.6% (21)	1.6% (24)	26.3% (23)
	59.4% (21)	85.3% (17)	2.8% (23)	3.8% (24)
	28.9% (23)	36.3% (23)	89.3% (15)	84.3% (20)
	64.5% (20)	85.8% (15)	48.3% (20)	81.7% (21)
	83.9% (15)	76.6% (19)	63.7% (19)	88.4% (19)
	87.5% (14)	82.9% (18)	95.6% (13)	95.5% (14)
	90.9% (13)	42.1% (22)	95.9% (11)	97.0% (10)
	98.8% (2)	71.3% (20)	38.9% (21)	97.9% (6)
	69.9% (19)	85.5% (16)	98.4% (2)	97.9% (5)

TABLE IV: Accuracy of selected classes. Numbers in parentheses denote the rank among all 24 classes. Even though only Austrian pictograms are shown in the table, all models are evaluated on the pictogram design indicated in the table header.

design they were trained on. Class 'Truck trailer' seems to cause most problems: DE-AT_c and DE-AT_n often confuse 'Truck trailer' with 'Truck'; AT_c-AT_n hardly ever confuses these two classes, but instead misclassifies 'Truck trailer' as 'Animal-drawn', such that in the total the accuracy of 'Truck trailer' drops as far as 1.1%, as can be seen in Table IV. It can also be seen that in all evaluations 'Motorcycle' is frequently misclassified as 'Truck', which might be owing to the three designs of 'Motorcycle' differing fairly strongly. An analogous statement applies to 'Power-driven'.

Table IV lists the per-class accuracy of the models for a couple of selected classes. Although the overall classification accuracy of all models drops considerably on foreign pictogram designs, there are blatant inter-class differences. In fact, a big deal of this drop is caused by only a few classes, namely those listed in Table IV; the others are correctly classified most of the time.

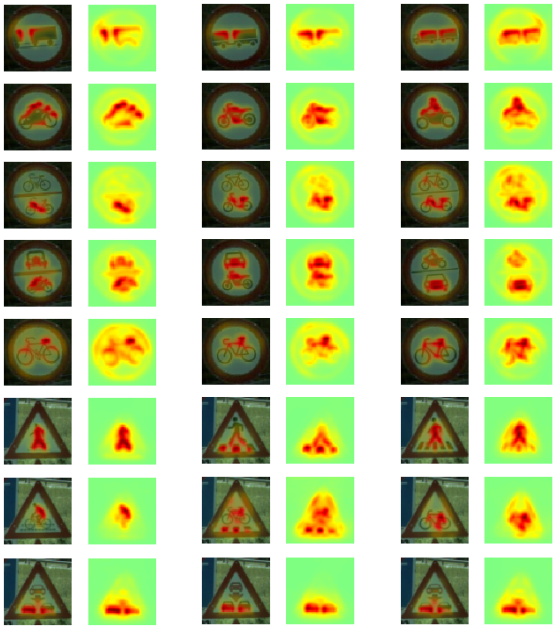
B. *Qualitative Explanations of the Models' Predictions*

We employed *layer-wise relevance propagation* (LRP) [30] for estimating the importance of image regions and -details to the classification models, in order to explain what information they base their predictions on. Among the multitude of possible parameter configurations of LRP we adhered to the one suggested for convolutional neural networks in [23] throughout.

Fig. 3 shows the average explanations of all correctly predicted test images of some selected classes, for the AT_c-AT_c, AT_n-AT_n and DE-DE experiments. Explanations are presented as heatmaps, where the color of a pixel indicates its relevance to the model. For each evaluation, the left column blends the heatmaps with the actual images to facilitate localization, whereas the right column only shows the heatmaps themselves. Evaluations CUR-AT_c and CUR-DE are spared since they exhibit a very similar relevance pattern as AT_c-AT_c and DE-DE, respectively.

As can be seen, all models strongly focus on the pictograms (or parts of them) when classifying a traffic sign image, and only sometimes take the border of the sign into account as well. On the one hand, this means that our models learned to pay attention to the 'right' details of an image and do not base their decisions on spurious artifacts in the background, and on the other hand, it means that the shape of the traffic signs does not really aid the models. This is not surprising, since the uniform circular and triangular shapes carry only little information for classifying the signs – especially if the pictograms alone are sufficient for that purpose. Only in some cases, where the pictograms of prohibitory and warning signs are similar in appearance, taking the shape into account can be beneficial. This phenomenon occurs, for example, with classes 'Cycle' and 'Cyclist crossing': for the models trained on current Austrian pictograms the shape of 'Cycle' seems to be quite important, whereas the other models pay more attention to the shape of 'Cyclist crossing'.

In classes 'Pedestrian crossing' and, to some extent, 'Cyclist crossing', the models trained on proposed new Austrian and German pictograms focus a lot on the zebra crossing



(a) AT_c-AT_c (b) AT_n-AT_n (c) DE-DE

Fig. 3: Average explanations of all correctly predicted test images of some selected classes. Reddish to yellowish hues indicate regions with evidence *in favor* of the predicted class and greenish hues indicate regions without any relevance.

at the bottom. The models trained on the current Austrian pictograms, on the other hand, completely ignore the (only barely visible) zebra crossing and instead focus on the person. This difference in attention might be one of the reasons why in AT_c-AT_n 'Cyclist crossing' and 'Pedestrian crossing' only achieve a comparatively low accuracy of 28.93% and 69.87%, respectively (cf. Table IV).

It can also be observed that sometimes the models only look at certain parts of the pictograms, and not at the whole pictograms. This effect is particularly visible in class 'Wrong way driver', where all models completely ignore the car at the top and almost entirely ignore the arrow as well. Apparently, the two cars at the bottom are sufficient for robustly distinguishing this class from the other 23 traffic sign classes in our experiments. Likewise, in class 'Cycle & Moped' the moped receives a lot more attention than the cycle, especially in AT_c-AT_c. Interestingly, in class 'Power-driven', whose pictogram is similarly split into an upper and a lower part, the relevance is distributed much more evenly across the car and the motorcycle.

Classes 'Truck trailer' and 'Motorcycle' allow us to speculate why the models fail to generalize to other pictogram designs in some cases. Namely, both classes differ between the three design groups in certain aspects the models pay a lot attention to. The fact that the truck in 'Truck trailer' is visible in its entirety in the German design seems to be important to the models, since quite some relevance is assigned to the front part of the truck. When comparing the two Austrian designs of this class one can also observe a subtle difference in the relevance pattern: only a small part of the truck is visible in the current Austrian design, leading to a vertical relevance pattern; in contrast, the proposed new Austrian design displays a slightly larger part of the truck, leading to a more horizontal pattern. Similarly, the fact that the current Austrian and German designs feature a person riding the motorcycle in class 'Motorcycle' seems to be important to the models. The proposed new Austrian design lacks a rider, which the models seem to compensate by paying more attention to the front wheel.

It must be noted, though, that further experiments are necessary to confirm the hypotheses expressed in the preceding paragraphs. The relevance patterns constructed by LRP or any other feature attribution method are only meant to illustrate which parts of an image are important to a model, but one must be careful when trying to draw conclusions why the model fails to classify some class correctly.

IV. DISCUSSION

The objective of our work was to answer three research questions regarding the machine readability of traffic signs:

1) Is there any significant difference between different pictogram designs (AT_c, AT_n, DE) in terms of machine readability?
2) Can traffic sign classification models trained on one pictogram design be safely deployed to traffic signs featuring a different design?
3) Can general 'design rules' for pictograms be formulated to improve machine readability?[3]

The first question can be answered readily: even though there *are* small differences in the observed model accuracies in AT_c-AT_c, AT_n-AT_n and DE-DE (cf. Table II), these differences are not significant. Hence, all three pictogram designs are equally well machine-readable.

The answer to the second question is also negative: if any of the models trained on one pictogram design is applied to a different design, its classification accuracy drops significantly, by about 15–23 percentage points. In this regard, it is particularly interesting to note that a few classes cause massive problems, whereas most of the others can still be classified accurately. Even more surprising is the fact that our models consistently generalize best between German and *current* Austrian pictograms (in both directions), although a human would probably find more similarities between German and *new* Austrian pictograms. We do not have any explanation for this phenomenon. Models trained on CUR generalize better to the unseen design AT_n than models trained on AT_c and DE alone. However, even here the performance drop of more than 13 percentage points, from 98.69% accuracy in CUR-CUR to 85.48% in CUR-AT_n, is significant. Hence, training on a more diverse set of pictogram designs leads to some improvement, but is still far from optimal. A larger study involving even more pictogram designs remains a possible subject for future research.

[3]Adding something like QR-codes would be an obvious affirmative answer, but in this work we focus on traffic infrastructure that can be processed by machines and humans alike.

Answering the third question is more intricate. Although we generated explanations for the models' predictions in Section III-B, formulating design rules for pictograms based on them is difficult. Still, what *can* be said is that deep neural networks perceive traffic signs differently than humans: humans try to *understand* the meaning of pictograms in order to classify them, machines only try to *distinguish* them. This is characteristic for discriminative methods, as it is exactly what they are meant to do. Distinguishing a fixed set of pictograms, however, might be possible based on small, semantically meaningless details. We hypothesize that this is the main reason why the models in our experiment fail to generalize to 'foreign' pictogram designs. Unfortunately, it is hardly possible to predict a-priori which details will be important to a classification model. The only general rule that can be formulated in this regard concerns the visibility of pictogram elements: the zebra crossings in classes 'Cyclist crossing' and 'Pedestrian crossing' of the current Austrian design consist of small, thin line segments that quickly become imperceptible when the images are corrupted, and hence the models do not pay attention to them. In the proposed new Austrian design the zebra crossings are far more pronounced and thus better visible, and the models *do* take them into account. Thin lines and overly small patches of ink should therefore be avoided.

Summarizing, the main takeaways of our work are as follows:

- Machines can handle different pictogram designs equally well, provided they have been trained on them.
- In the realistic scenario that an ADAS should correctly recognize traffics signs with different pictogram designs, the models must be trained appropriately. This can be achieved by either training one single classifier on a data set encompassing many different designs or by training a separate classifier for each design.
- If existing pictograms are replaced by a new design, classification models will likely have to be updated. Since acquiring large real-world data sets is time-consuming and only possible once the actual traffic signs have been replaced, it might be necessary to resort to synthetic data sets as presented in this paper, instead.

V. Limitations and Future Work

When defining our experimental setup, we had to fix certain parameter values that are up to discussion and could be revised in future extensions of our experiments. First, we only considered 24 traffic sign classes from categories 'prohibitory' and 'warning'. Actual classifiers deployed in ADAS must be trained on a much wider variety of classes and may hence exhibit a different behavior w. r. t. sensitivity to pictogram design, frequently confused classes, and attention patterns. Still, we believe that our reduced setting approximates reality sufficiently well for making our findings hold more generally. A similar statement applies to the investigated model architectures. Extending the experiments to more architectures, like Vision Transformers [13], for obtaining more reliable results is certainly possible. Furthermore, traffic sign recognition systems deployed in ADAS are not disclosed to the scientific community, so one could question whether our findings are even applicable to them. Indeed, we merely want to encourage developers of ADAS to consider our experimental results and, if appropriate, conduct similar experiments with their own traffic sign classifiers. Yet, we think that the highly similar results of two fairly distinct architectures present strong evidence that our findings are not limited to the concrete architectures under consideration.

Another point of discussion concerns the image corruption strategy. From the vast space of conceivable corruption methods we picked some that we deemed either realistic or particularly interesting, but many others would have been at our disposal, too. In future experiments, one could in particular try to incorporate corruptions that are specific to traffic signs, like some kind of 'over-exposure' where, due to the production process and reflectivity of the traffic sign foil, brighter areas seem to 'grow' and hide parts of darker neighboring areas, making small and fine pictogram elements seemingly disappear. Furthermore, we focused on simulating distance by spatially downsampling the images at varying degrees, but we did not apply other geometric transformations like rotations and perspective distortions. In addition to the degree of downsampling, one could systematically vary the intensities of the 'secondary' corruptions (rain, noise, blur, etc.) as well. A complementary augmentation strategy could specifically target the pictograms themselves, for instance, by systematically (re)moving vertices in vectorial versions of the pictograms.

As discussed above, the models we obtained are not very robust w. r. t. 'foreign' pictogram designs. One way to counter this could be forcing the models to pay more attention to the global shape of the pictograms, instead of small details. This, in turn, can perhaps be achieved by borrowing ideas from current research on *adversarial attacks* [36], [17], like *adversarial training* [26], [38]. Alternatively, one could also try to preprocess the images before training and applying a model, by applying a low-pass- or bilateral filter that destroys high-frequency information and thereby biases the model towards low-frequency shape information. Repeating our experiments with adversarially trained models or said input preprocessing could be an interesting direction for future research.

Finally, it would be interesting to see how well the models trained on our purely synthetic data would perform on real-world data, like GTSRB. This could serve as sort of a 'sanity check' to ensure that the synthetic data sets resemble reality sufficiently well. Of the 24 classes considered in our experiments only seven are included in GTSRB, though, rendering an exhaustive evaluation impossible.

Acknowledgment

We thank Stefan Egger for providing us with the proposed new Austrian pictogram designs and for his suggestions regarding our experimental setup. We also thank Isabell Ganitzer for carefully proofreading a draft version of this paper, and the anonymous reviewers for their valuable comments.

REFERENCES

[1] "VzKat 2017," http://www.vzkat.de, 2017, online.

[2] "AI.Reverie," https://aireverie.com/, 2021, online.

[3] "Anyverse," https://anyverse.ai/, 2021, online.

[4] "Cognata Traffic Sign Datasets," https://www.cognata.com/traffic-sign-datasets/, 2021, online.

[5] "CVEDIA," https://www.cvedia.com/, 2021, online.

[6] Austrian Federal Ministry for Digital and Economic Affairs, "Straßenverkehrsordnung 1960, Fassung vom 12.11.2018," https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10011336, §§ 50 and 52, 2018, online.

[7] A. Azulay and Y. Weiss, "Why do deep convolutional networks generalize so poorly to small image transformations?" *Journal of Machine Learning Research*, vol. 20, no. 184, pp. 1–25, 2019.

[8] R. Belaroussi, P. Foucher, J.-P. Tarel, B. Soheilian, P. Charbonnier, and N. Paparoditis, "Road sign detection in images: A case study," in *International Conference on Pattern Recognition*, 2010, pp. 484–488.

[9] C. Berghoff, P. Bielik, M. Neu, P. Tsankov, and A. von Twickel, "Robustness testing of AI systems: A case study for traffic sign recognition," in *Artificial Intelligence Applications and Innovations*, ser. IFIP Advances in Information and Communication Technology, I. Maglogiannis, J. Macintyre, and L. Iliadis, Eds. Springer, 2021, pp. 256–267.

[10] A. v. Bernuth, G. Volk, and O. Bringmann, "Simulating photo-realistic snow and fog on existing images for enhanced CNN training and evaluation," in *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2019, pp. 41–46.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[12] S. Dodge and L. Karam, "Understanding how image quality affects deep neural networks," in *International Conference on Quality of Multimedia Experience (QoMEX)*, 2016, pp. 1–6.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2020.

[14] C. Ertler, J. Mislej, T. Ollmann, L. Porzi, G. Neuhold, and Y. Kuang, "The mapillary traffic sign dataset for detection and classification on a global scale," in *Computer Vision – ECCV 2020*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Springer, 2020, pp. 68–84.

[15] R. Geirhos, C. R. M. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann, "Generalisation in humans and deep neural networks," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 7538–7550.

[16] C. Gámez Serna and Y. Ruichek, "Classification of traffic signs: The european dataset," *IEEE Access*, vol. 6, pp. 78 136–78 148, 2018, conference Name: IEEE Access.

[17] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations (ICLR)*, 2015.

[18] C. Grigorescu and N. Petkov, "Distance sets for shape filters and shape recognition," *IEEE Transactions on Image Processing*, vol. 12, no. 10, pp. 1274–1286, 2003.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[20] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *International Conference on Learning Representations (ICLR)*, 2019.

[21] A. B. Jung, K. Wada, J. Crall, S. Tanaka, J. Graving, C. Reinders, S. Yadav, J. Banerjee, G. Vecsei, A. Kraft, Z. Rui, J. Borovec, C. Vallentin, S. Zhydenko, K. Pfeiffer, B. Cook, I. Fernández, F.-M. De Rainville, C.-H. Weng, A. Ayala-Acevedo, R. Meudec, M. Laporte, *et al.*, "imgaug," 2020. [Online]. Available: https://github.com/aleju/imgaug

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[23] M. Kohlbrenner, A. Bauer, S. Nakajima, A. Binder, W. Samek, and S. Lapuschkin, "Towards best practice in explaining neural network decisions with lrp," in *International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–7.

[24] F. Larsson and M. Felsberg, "Using fourier descriptors and spatial models for traffic sign recognition," in *Image Analysis*, ser. Lecture Notes in Computer Science, A. Heyden and F. Kahl, Eds. Springer, 2011, pp. 238–249.

[25] J. Li and Z. Wang, "Real-time traffic sign recognition based on efficient CNNs in the wild," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 3, pp. 975–984, 2019.

[26] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations (ICLR)*, 2018.

[27] A. Maletzky, S. Thumfart, and C. Wruß, "Comparing the machine readability of traffic sign pictograms in Austria and Germany," *arXiv:2109.02362 [cs.CV]*, 2021.

[28] C. Michaelis, B. Mitzkus, R. Geirhos, E. Rusak, O. Bringmann, A. S. Ecker, M. Bethge, and W. Brendel, "Benchmarking robustness in object detection: Autonomous driving when winter is coming," *arXiv:1907.07484 [cs, stat]*, 2019.

[29] A. Møgelmose, M. M. Trivedi, and T. B. Moeslund, "Vision based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 4, pp. 1484–1497, 2012.

[30] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-wise relevance propagation: An overview," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, ser. Lecture Notes in Computer Science, vol. 11700, 2019, pp. 193–209.

[31] A. L. Pavlov, P. A. Karpyshev, G. V. Ovchinnikov, I. V. Oseledets, and D. Tsetserukou, "IceVisionSet: lossless video dataset collected on russian winter roads with traffic sign annotations," in *International Conference on Robotics and Automation (ICRA)*, 2019, pp. 9597–9602, ISSN: 2577-087X.

[32] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3234–3243.

[33] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *International Journal of Computer Vision*, vol. 126, no. 9, pp. 973–992, 2018.

[34] C. Sitawarin, A. N. Bhagoji, A. Mosenia, M. Chiang, and P. Mittal, "DARTS: Deceiving autonomous cars with toxic signs," *arXiv:1802.06430 [cs]*, 2018.

[35] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition," *Neural Networks*, vol. 32, pp. 323–332, 2012.

[36] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations (ICLR)*, 2014.

[37] D. Tabernik and D. Skočaj, "Deep learning for large-scale traffic-sign detection and recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 4, pp. 1427–1440, 2020.

[38] J. Tack, S. Yu, J. Jeong, M. Kim, S. J. Hwang, and J. Shin, "Consistency regularization for adversarial robustness," *arXiv:2103.04623 [cs]*, 2021.

[39] D. Temel, M.-H. Chen, and G. AlRegib, "Traffic sign detection under challenging conditions: A deeper look into performance variations and spectral characteristics," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–11, 2019.

[40] R. Timofte, K. Zimmermann, and L. Van Gool, "Multi-view traffic sign detection, recognition, and 3d localisation," *Machine Vision and Applications*, vol. 25, no. 3, pp. 633–647, 2014.

[41] G. Volk, S. Müller, A. v. Bernuth, D. Hospach, and O. Bringmann, "Towards robust CNN-based object detection through augmentation with synthetic rain variations," in *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2019, pp. 285–292.

[42] Y. Yang, H. Luo, H. Xu, and F. Wu, "Towards real-time traffic sign detection and classification," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 7, pp. 2022–2031, 2016.

[43] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, "Traffic-sign detection and classification in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2110–2118.

Efficient Instance Segmentation of Panoramic Images of Indoor Scenes

Werner Bailer and Hannes Fassold

Abstract— **This paper addresses the issue of efficient 2D instance segmentation of 360° images of indoor scenes. In particular, we study the use of equirectangular convolutions and the impact of different approaches to handle wrap-around areas. We consider the use of Mollweide projection as a representation for performing segmentation, and we provide a toolchain to prepare the Matterport panoramic images for use in workflows designed for COCO-style annotated datasets. The results show no significant differences between using regular and equirectangular convolutions. While the Mollweide projection allows for segmentation of otherwise missed objects, the overall results do not outperform analysis on equirectangular projection.**

I. INTRODUCTION

In many application areas (e.g., interior design, furniture retailing or renovation), communication with a customer or future user during the planning and design phase is crucial to select the right products and configurations. Making this communication process effective saves costs, avoids later modifications, and results in providing tailored solutions and higher customer satisfaction. Augmented Reality (AR) has the potential to make these communication processes highly effective and provide a better experience for the customer. However, AR content needs to be created by experts from the respective domains, who often lack IT and media skills, and shall provide a lightweight AR experience for the customer. Current AR authoring solutions are quite complex and require manually creating scenes or rely on objects prepared with even more complex applications (e.g. CAD). In order to facilitate this process, a simple capture process (e.g., using consumer grade 360° cameras) and intelligent scene understanding tools are needed.

One important component is segmenting and classifying the relevant objects such as furniture in interior scenes. In particular, we aim to perform instance segmentation for indoor scenes in single panoramas of rooms. This shall also be possible on consumer hardware with limited processing capabilities. In order to process the 360° images, we aim to avoid training or fine-tuning models specifically for 360° data. This is motivated by the fact that annotated datasets for object segmentation on panoramic images are very scarce. Due to the efficiency requirements, performing the analysis on separate viewports of the 360° image is not feasible.

The authors are with DIGITAL – Institute for Information and Communication Technologies at JOANNEUM RESEARCH, Graz, Austria, {firstname.lastname}@joanneum.at

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 951900 ATLANTIS ("Authoring tool for indoor augmented and diminished reality experiences"). The authors thank Nikolaos Zioulis from CERTH ITI for improving the Matterport data preparation script, and Hermann Fürntratt for help with the COCO annotation script.

The contributions of this paper are: (i) we study the use of equirectangular convolutions and the impact of handling the wrap-around areas, (ii) we consider the use of Mollweide projection as a representation for performing the segmentation, and (iii) we provide a toolchain to prepare the Matterport panoramic images for use in workflows designed for COCO-style annotated datasets.

The rest of this paper is organized as follows. Section II discusses related work and Section III presents the approaches that were investigated. Section IV discusses the evaluation (including dataset preparation) and the obtained results, and Section V concludes the paper.

II. RELATED WORK

Impressive progress has been made in instance segmentation of indoor environments represented as point clouds. Such point clouds can be obtained from capturing the scene with multiple views or depth sensors. Recent approaches such as PointGroup [16], 3D-SIS [14] and 3D-MPA [10] show good performance on benchmarks such as ScanNet [7]. However, in many consumer application scenarios, depth information is not available, and thus 2D approaches are required.

We are thus interested in an efficient and reliable 2D instance segmentation approach. A well known approach is Mask R-CNN [12], a two stage instance segmentation based on Faster R-CNN. Masklab [5] is a further evolution of this type of approaches. In terms of efficiency, single stage approaches are preferable. Recent methods showing good performance on benchmark datasets include SOLO v2 [24], Yolact++ [2], proposal free instance segmentation [15] and SipMask [3].

We aim to apply instance segmentation to 360° images. Different approaches to handle this issue have been proposed in literature. One group of methods requires specific training on 360° images or at least fine-tuning. This can be done by adapting early layers of a pretrained network to work on equirectangular images, which is proposed in [22] and tested for object detection using VGG and Faster R-CNN. [6] follow a similar approach with SphereNet, learning a network adjusted to equirectangular inputs. The use of icosahedral Snyder equal-area (ISEA) projections is proposed in [8] and results in significant improvement for semantic indoor segmentation on the SUMO dataset.

To avoid the need for specifically training the network on panoramic data, [25] perform segmentation on multiple stereographic projections. Equirectangular convolutions are proposed in [23] as a convolution kernel for equirectangular images that adjusts the input values to the image positions,

including handling of wrap-around. A similar approach is proposed in [9], with generalised convolutions that use a mapping function. That paper analyses different mapping functions and proposes mapping to a geodesic grid. Equirectangular convolutions have been recently used for indoor semantic segmentation [11], and the authors report small improvements over processing the equirectangular image with standard convolutions.

While a number of approaches for handling 360° images in CNNs have been proposed, many of them require some kind of training or fine-tuning, which limits the practical application. Using specific types of convolutions is reported to improve performance (at least slightly) in some papers, but most of the work deals with object detection rather than segmentation. We are thus interested to study the impact of these choices in our applications, as well as the use of the Mollweide projection, which to the best of our knowledge has not yet been investigated for this purpose.

III. Investigated approaches

Due to the lightweight implementation and the potential to run the method also on a mobile device, we select Yolact++ [2] as the basis of our work. We implement two approaches for processing 360° images with Yolact++: the first one is to integrate equirectangular convolutions, and the second is to transform input images using the Mollweide projection.

For both equirectangular and Mollweide projected images we also optionally extend the image to wrap-around the seam of the panorama in order to facilitate segmentation of objects cut across by the seam. We found experimentally that using $1/8$ of the image width is a useful value for indoor scenes to ensure that objects of interest become visible in an unseparated way on at least one side.

A. Equirectangular content processing

Yolact++ uses ResNet-101 [13] with FPN [19] as its backbone. We thus replace the convolutions in the first layer of the backbone network with the equirectangular convolutions proposed in [23], leaving the parameters of the convolutions otherwise unchanged. In particular, we use the EquiConv Pytorch implementation[1]. These convolutions change which pixels are used as input depending on the position, simulating regular sampling on a spherical surface. This includes handling wrap-around, i.e., accessing pixels from the opposite image border when necessary.

While EquiConv is only used in one layer, the runtime difference in inference is still noticable, compared to the highly optimized implementations for regular convolutions, which are increasingly available (including on mobile devices).

B. Mollweide projection

The Mollweide projection [17] is a pseudocylindrical, equal-area projection. It is also known as homolographic projection or elliptical projection. In contrast to the equirectangular projection, it does not stretch areas near the poles. As a downside, the Mollweide projection bends vertical longitude lines, whereas the equirectangular projection keeps them straight. So each projection has its strong points as well as weak points. In order to retain at least to a certain degree the desirable properties of both projections, we propose a mix of both Mollweide and equirectangular projection, which we will term *hybrid Mollweide projection* in the following. We define a blending factor α in the range $[0,1]$, which allows use to interpolate smoothly between the two projections. We retrieve the standard Mollweide projection by setting $\alpha = 0.0$, the equirectangular projection by setting $\alpha = 1.0$ and a mix where both projections are weighted equally by setting $\alpha = 0.5$. The implementation of standard Mollweide and hybrid Mollweide projections follows the equations given in [26] for the equirectangular projection, with a few modifications in some places. Specifically, the equations for the conversion between the sampling point (u,v) and longitude-latitude (ϕ,θ) have to be modified properly in the following way: Equation (6) from [26] is to be replaced by

$$u = (x + 0.5) \,/\, W'$$

with

$$W' = ((1-\alpha)\; d(\theta) + \alpha) \cdot W$$

$$d(\theta) = \sqrt{1 - \left(\frac{2}{\pi}\,\theta\right)^2}$$

Another point we have to take into account is that the longitude ϕ is cyclic, meaning that the image pixels on the left and right border of the Mollweide projection actually belong to the same region on the sphere. To address this, we add additional border pixels in each image row, on the left and right side. The border pixels are taken from the respective inner region of the other side (so the border pixels added on the left side are taken from the inner region of the right side, and vice versa). Figure 1 shows examples of (hybrid) Mollweide projections.

IV. Evaluation

A. Dataset preparation

For evaluation we require a dataset that provides natural panoramic images of indoor scenes. A number of the indoor datasets containing panoramic images, such as InteriorNet [18] and Structured3D [27], contain only synthetic images. Sun-CG [21] (and the derived SUMO dataset) were very actively used datasets for this purpose, but the dataset has been withdrawn. Thus there are two remaining datasets that meet this condition: Matterport3D [4] and 2D-3D-S [1]. As Matterport3D contains rather private homes than office spaces, we selected this dataset. The dataset contains 10,800 panoramic views of 90 houses. As we use a model trained on the COCO dataset, we only use the test split of Matterport3D, consisting of 18 houses with 1,848 panoramic views.

While panoramic RGB images are provided with the dataset, the instance and semantic segmentation ground truth maps are not. The scenes have been labelled on 3D meshes, and thus the annotations are provided in this format.

[1] https://github.com/palver7/EquiConvPytorch

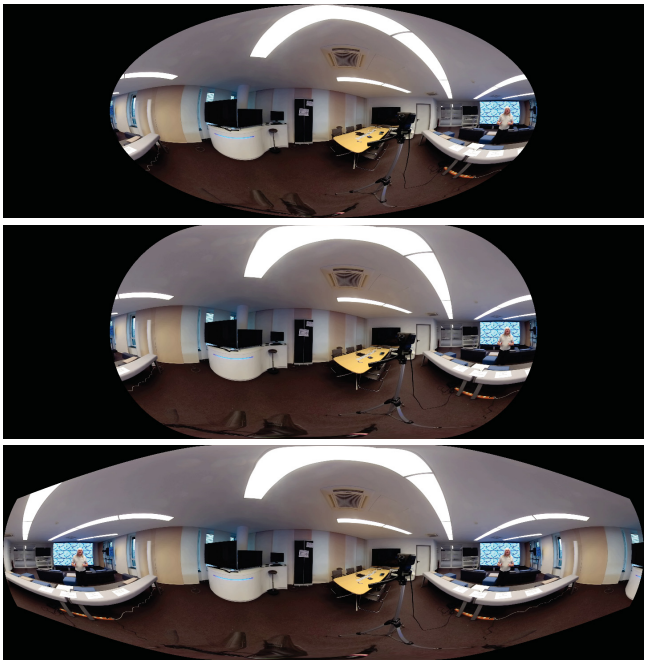


Fig. 1. Example of Mollweide projection (top), hybrid Mollweide projection ($\alpha = 0.5$, middle) and hybrid Mollweide projection with border ($\alpha = 0.5$, border=0.125, bottom).

We modified the mpview tool, that is provided with the dataset[2], in order to batch render the semantic and instance segmentation maps corresponding to each view. As support for 360° cameras is not easy to integrate into this viewer, we generate the segmentation maps for each of the 18 tiles used to compose the panoramas in the dataset, and perform the stitching process for the segmentation maps. Apart from some mislabelled parts of the mesh, some object and wall meshes have holes, that makes other objects visible (e.g., a TV screen from the outside view of a house). These issues, that cannot be resolved automatically, create some level of noise in the annotations which we have to accept due to lack of resources to manually fix them.

The annotations are provided for a set of 40 indoor classes specific for this dataset. These classes mostly (though not fully) overlap with the more commonly used NYU40 set of classes [20]. In order to work with models pretrained on the COCO dataset, we use the overlapping set of classes between Matterport3D and COCO: chair, couch, potted plant, bed, dining table, toilet, TV, sink.

Most semantic and instance segmentation methods support the COCO annotation format. We have thus created a tool to convert the Matterport3D segmentation maps to COCO annotations. This involves generating polygons from the object masks, for which we use pycococreator[3]. The COCO annotation format does not support the notion of subtracting partial polygons, thus we apply hole filling to the binary mask. In order to reduce the issue of border pixels or small regions caused by triangles cutting through the surface of other objects, we also apply morphologic closing. However, this does in many cases not remove the false annotations caused by holes in the mesh mentioned above.

One other property of the Matterport dataset is that many of the rooms are rather "loft-style", i.e., other capture locations are visible in the background. Most objects thus appear multiple times, once quite prominently in the room being captured, and one or more times in another (part of the) room. This also results in a large number of small annotated objects. In fact, 64.8% of the object instances are smaller than 0.05% of the image area, and 85.1% of the object instances occur more than once. The size differences are significant: in 62.8% of cases the smallest occurrence has an area of 1% or less than the largest occurrence of the same object instance.

For equirectangular images, annotations extending across the seam of the image will result in polygons at the left and right borders of the image. For the cases where borders for handling wrap-around have been added (either to equirectangular, Mollweide or hybrid images), we process the annotations in the border regions to keep only those that continue from the image center into the border, but remove those that only start in the border regions (and are likely to wrap around, unless they are small).

Our toolchain for preparing the Matterport3D dataset is made available at `https://github.com/atlantis-ar/matterport_utils`. It consists of a modified version of the mpview tool for generating class and instance segmentation maps, a script of combining source images and generated maps into panoramas, and a script for creating COCO style annotation files.

B. RESULTS

We compare the performance of a Yolact++ model trained on COCO applied to the panoramic Matterport3D views under different conditions in terms of projection, convolution type and wrap-around handling. We measure the average precision (AP) of detected masks at overlaps (IoU) of 50% (AP@50) and 70% (AP@70). An overview of the results is provided in Table I. In order to show the impact of the many small regions from far away objects, we provide results for evaluating against the unfiltered ground truth as well as against a ground truth where objects smaller than 0.5% of the image area have been filtered out. Note that this is a very conservative choice, that will not remove all the multiply depicted objects, but has been chosen to ensure that no smaller foreground objects are removed. To put the results in relation, it is worth noting that the current state of the art for 2D instance segmentation on the ScanNet benchmark[4] is 0.358 in terms of AP@50 (consisting of regular rather than panoramic images).

From the results we can see that there is no significant difference between using regular convolutions and equirectangular convolutions. Also the configurations adding extra

[2] `https://github.com/niessner/Matterport/tree/master/code/gaps/apps/mpview`

[3] `https://github.com/waspinator/pycococreator`

[4] `http://kaldir.vc.in.tum.de/scannet_benchmark/semantic_instance_2d.php?metric=ap50`

Projection	conv	min	wrap	border	AP	
					IoU50	IoU70
Equirect	regular	0.0	no	0	15.07	6.64
Equirect	regular	0.5	no	0	26.88	12.77
Equirect	equiconv	0.5	yes	1/8	26.82	12.49
Equirect	regular	0.5	no	1/8	25.21	11.86
Equirect	regular	0.5	yes	1/8	24.92	11.78
Mollweide $\alpha = 0.0$	regular	0.5	no	1/8	16.10	6.63
Mollweide $\alpha = 0.0$	regular	0.5	yes	1/8	15.97	6.58
Mollweide $\alpha = 0.5$	regular	0.5	no	1/8	21.95	10.10
Mollweide $\alpha = 0.5$	regular	0.5	yes	1/8	21.82	9.93

TABLE I

OVERVIEW OF THE RESULTS OBTAINED WITH A YOLACT++ MODEL TRAINED ON COCO. *conv* REFERS TO TYPE OF CONVOLUTION USED, *min* DESCRIBES THE MINIMUM AREA OF OBJECTS (IN PERCENT OF THE IMAGE AREA) THAT WERE RETAINED IN THE GROUND TRUTH, *wrap* DESCRIBES WHETHER WRAP AROUND HANDLING HAS BEEN APPLIED TO THE GROUND TRUTH AND *border* SPECIFIES THE WIDTH OF A BORDER BEING ADDED.

borders for wrap around handling perform very similar, though slightly worse. In addition, filtering the ground truth to have each object in the border region only once performs slightly worse (for all projections and overlaps) than not doing so. The reasons seem to be that it is not necessarily the more prominent version of the object that is better segmented, and that partial objects at borders are sometimes quite well segmented. The pure Mollweide projection performs clearly worse, and the results improve when we mix the projections.

Figure 2 shows an example of the results obtained with the different configurations and the detection of some objects, e.g., the second chair, the falsely detected TV in the office and the bed visible from the room next door. While the Mollweide projection performs generally worse than equirectangular, there are some objects that are only detected in Mollweide projection, and get lost already in the hybrid projection. We also observe some differences between the equirectangular projection with and without border. The presumption is that the aspect ratio change due to adding the border also plays a role in this behaviour.

We have performed further experiments with the SOLOv2 [24] framework, training it on the COCO and ScanNet datasets. The results indicate that the small differences in terms of performance between regular or equirectangular convolutions also hold for other models and datasets. However, the dataset determines how well the model generalises to equirectangular images. We observe that models trained on COCO images generally provide better segmentation quality (in particular, concerning the accuracy of the mask) for equirectangular images than those trained on ScanNet.

V. CONCLUSION

In this paper we have studied the problem of efficient 2D instance segmentation of 360° images of indoor scenes. We have analysed different ways of preparing equirectangular content, and assessed the use of regular vs. equirectangular convolutions on equirectangular projections. In addition, we consider the Mollweide projection as an alternative projection. We performed evaluation for the different configurations on panoramic images from the Matterport3D dataset. One contribution of this paper is thus a toolchain for preparing the panoramic dataset, and provide class and instance labels in COCO-style annotation format for use with a wide range of object detection and segmentation methods.

The conclusion from our experiments is that using equirectangular convolutions does not improve performance, but is computationally less efficient than the well optimised implementations for regular convolutions. While the Mollweide projection allows for segmentation of otherwise missed objects in a number of cases, the overall results do not outperform those on equirectangular projection. It needs to be further studied, if combining results from different projections provides benefits and justifies the increased computational effort.

REFERENCES

[1] I. Armeni, S. Sax, A. Zamir, and S. Savarese, "Joint 2d-3d-semantic data for indoor scene understanding," *ArXiv*, vol. abs/1702.01105, 2017.

[2] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact++: Better real-time instance segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[3] J. Cao, R. M. Anwer, H. Cholakkal, F. S. Khan, Y. Pang, and L. Shao, "Sipmask: Spatial information preservation for fast image and video instance segmentation," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 1–18.

[4] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," *International Conference on 3D Vision (3DV)*, 2017.

[5] L.-C. Chen, A. Hermans, G. Papandreou, F. Schroff, P. Wang, and H. Adam, "Masklab: Instance segmentation by refining object detection with semantic and direction features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4013–4022.

[6] B. Coors, A. Paul Condurache, and A. Geiger, "Spherenet: Learning spherical representations for detection and classification in omnidirectional images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 518–533.

[7] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[8] M. Eder and J.-M. Frahm, "Convolutions on spherical images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 1–5.

[9] M. Eder, T. Price, T. Vu, A. Bapat, and J.-M. Frahm, "Mapped convolutions," arXiv:1906.11096, Tech. Rep., 2019.

[10] F. Engelmann, M. Bokeloh, A. Fathi, B. Leibe, and M. Nießner, "3d-mpa: Multi-proposal aggregation for 3d semantic instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9031–9040.

[11] J. Guerrero-Viu, C. Fernandez-Labrador, C. Demonceaux, and J. J. Guerrero, "Whats in my room? object recognition on indoor panoramic images," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 567–573.

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

Fig. 2. Example from Matterport3D dataset: input image (upper left), results on equirectangular projection with standard convolution (upper right) / equiconv (middle left), with wraparound (middle right), Mollweide projection with standard convolution (lower left: $\alpha = 0.0$, lower right: $\alpha = 0.5$). Best viewed in color.

[14] J. Hou, A. Dai, and M. Nießner, "3d-sis: 3d semantic instance segmentation of rgb-d scans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4421–4430.

[15] Y.-C. Hsu, Z. Xu, Z. Kira, and J. Huang, "Learning to cluster for proposal-free instance segmentation," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–8.

[16] L. Jiang, H. Zhao, S. Shi, S. Liu, C.-W. Fu, and J. Jia, "Pointgroup: Dual-set point grouping for 3d instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4867–4876.

[17] M. Kennedy and S. Kopp, *Understanding map projections*. Redlands, California: Esri Press, 2011.

[18] W. Li, S. Saeedi, J. McCormac, R. Clark, D. Tzoumanikas, Q. Ye, Y. Huang, R. Tang, and S. Leutenegger, "Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset," in *British Machine Vision Conference (BMVC)*, 2018.

[19] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[20] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.

[21] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1746–1754.

[22] Y.-C. Su and K. Grauman, "Learning spherical convolution for fast features from 360 imagery," in *Advances in Neural Information Processing Systems*, 2017, pp. 529–539.

[23] K. Tateno, N. Navab, and F. Tombari, "Distortion-aware convolutional filters for dense prediction in panoramic images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 707–722.

[24] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "Solov2: Dynamic and fast instance segmentation," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[25] W. Yang, Y. Qian, J.-K. Kämäräinen, F. Cricri, and L. Fan, "Object detection in equirectangular panorama," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 2190–2195.

[26] Y. Ye, E. Alshina, and J. Boyce, "Algorithm descriptions of projection format conversion and video quality metrics in 360lib," Joint Video Exploration Team (JVET), Tech. Rep., 2017.

[27] J. Zheng, J. Zhang, J. Li, R. Tang, S. Gao, and Z. Zhou, "Structured3d: A large photo-realistic dataset for structured 3d modeling," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 519–535.

High-Speed Stereoscopic Fragment Tracking in Industrial Filter Cleaning

Friedrich Holzinger[1], Michael Schneeberger[2], Manfred Klopschitz[2], Martina Uray[2], Matthias Rüther[1] and Gernot Krammer[1]

***Abstract*— Dust filtration is a critical operation in industrial processes, especially for environmental gas cleaning. Thereby, contaminated gas is passed through a fabric that holds back solid particles. Over time, these particles form a compact layer on the fabric surface further denoted as filter cake. This filter cake is periodically cleaned off via reverse air jet pulses, which cause an explosive breakup of the cake. To gather insights on the mechanics of this breakup, a high-speed, contactless 3D-monitoring method is presented and evaluated on laboratory scale. Results show that, despite the untextured appearance of the filter cake, for the first time a continuous three-dimensional evaluation of the filter cake motion is made possible. From our experiments, a novel data set is provided which will allow the derivation of mechanical parameters for filter cake fragments and allow the implementation and validation of dynamic simulations further on.**

I. INTRODUCTION

Industrial filter systems typically are implemented as bag-houses, i.e. batteries of hanging, cylindrical filter bodies which receive contaminated gas on the outside and deliver cleaned gas on the inside of each cylinder. They are applied in processes like waste compustion to reduce the emission of solid particles. In this process a permeable filter medium in the gas stream blocks particles up to $10\times$ smaller than its pore diameter, from passing through. As a consequence, dust particles deposit on the filter surface and form a powdery layer, which is commonly denoted as filter cake. Over time, the layer thickness increases and with it also the flow resistance. To ensure a continuous process, it is mandatory to remove the filter cake after a certain time or pressure limit. This is often accomplished by a reverse jet pulse of compressed air of about 100 milliseconds, which is applied in opposite gas flow direction and pushes the filter cake off the filter surface. It is common practice that the pulse is applied in-situ, without the interruption of filtration. This combination of a continuous forward air flow and a reverse air pulse creates a complex breakup scenario of filter cake with fragments partially flying off the filter surface, dispersing or being reattached. In this work, filter cleaning scenarios are reproduced at laboratory scale and a computer vision system is presented, which allows to observe the cleaning process at high temporal resolution. A considerable challenge is the 3D tracking of the decomposing filter cake components, despite the white, untextured surface, the high speed of motion and the continuous decomposition of the components. We thereby introduce a semi-automated approach, where a high-speed stereoscopic camera-projector-system is used to generate a sequence of depth maps, followed by manual annotation of particle agglomerates.

*This work was supported by the Austrian Science Fund (FWF) grant no. P30447-N32. Thanks to the Institute of Computer Graphics and Vision of the TUG for providing parts of the equipment.

[1]TUG - Institute of Process and Particle Engineering friedrich.holzinger@student.tugraz.at, krammer@tugraz.at

[2]JOANNEUM RESEARCH Forschungsgesellschaft mbH - DIGITAL name.surname@joanneum.at

A. Background of Optical Instruments in the Field of Filtration

In releated work, 2D image sensors were used to observe the partial removal of dust from filters [6] (p. 213), [4]. Usually, one camera scanned one section after another and 3D data allowed a quasi-static layer height analysis. This data was obtained with a single camera e.g., by directing structured light patterns on the surface [2]. Other groups demonstrated the benefits of using a stereo camera among other imaging techniques [8], [7] as usually perfect alignment of the equipment is not required, and a single shot is recorded in a short period of time enabling real-time applications. In this work the high-speed stereo camera provides time-resolved 3D images for above analysis of the course of patches that liberated from the filter during the cleaning pulse.

II. OPTICAL SYSTEM INTEGRATION

In order to reproduce a filter cleaning process at laboratory scale, first the minimum required filter area must be determined, which has to be used in a lab test in order to produce representative results. In a production environment, industrial filter material is always spanned over a grid of support rods. The grid size ultimately determines the size of a laboratory setup, because it defines an isolated cell of filter operation and a scallopping of the filter material is confined by the supporting rods. This grid size is reported to be in the range of 30 to 40 *mm* in width and 200 to 400 *mm* in length. As the scalloping effect is evident circumferentially, the width of the grid is the limiting dimension for lab scale reproduction. At the lab-scale rig, this dimension is reflected in the 70 *mm* diameter of the filter fixation ring, which is inserted between a cylinder and a pressing ring. The test rig is designed to operate at normative test conditions standardized in ISO 11057, therefore the ring orifice makes 56 *mm* of the filter visible. Particulate gas is fed directly towards the filter via a diffusor outlet. Downstream, the cylinder is connected to a blower that ensures a forced direction of flow. The

dust-containing volume between raw gas outlet and filter is confined by a metallic chamber that allows the obeservation by an optical system from the outside. The construction dimensions are shown on the sketch on the left half of Figure 1. A side channel is used to further reduce the minimum gas flow rate, resulting in a filtration speed of about 20 m/min. The intended measurement volume of the vision system is given by a cylinder with the filter surface at the bottom and a height of 20 *mm*. The cameras are situated opposite the filter, where a 6 *mm* thick composite glass replaces the metallic wall to allow optical access and still prevent the optical system from dust deposits. The glass can be cleaned by a magnetic wiper, inside from the outside. As shown in the photograph on the right half of Figure 1, the optical system consists of two synchronized cameras, capturing images in 512 *px* x 600 *px* resolution at a framerate of 500 Hz. Due to the high motion speed, the camera exposure is set to a maximum of 750 μs and synchronized to a high-power LED flashlight. To assist an automated dense stereo reconstruction, the directed LED illumination is enhanced by a static random dot pattern, generated by a Laser light source. In this combination, the directed LED light generally generates crisp shadows at filter cake cracks and edges, while the random dot pattern generates a dense surface texture. In this way, the conditions at the onset of the cleaning pulse can be examined, which reaches a steady state after about 20 *ms*. The acquisition of an image sequence is triggered by the pressure valve of the reverse jet pulse. The optical distance of the cameras, both focused on the center point of the top filter surface, is about 374 *mm* – each with a stereo angle of about 15° giving a stereo base (the distance of the optical lenses) of about 100 *mm*.

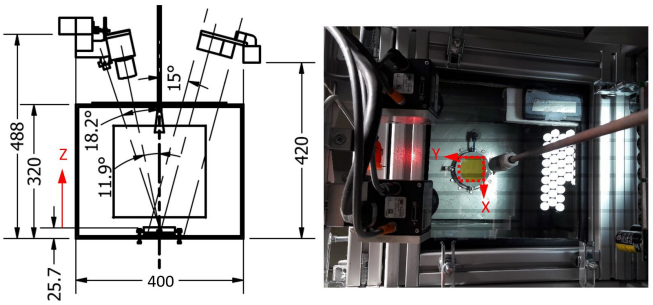


Fig. 1. Left: Sketch of test rig in front view with outlined metallic box. Right: Photograph of test rig from top. From left to right: pattern generator, stereo camera, filter fixed in holder, powder feed with diffusor, magnetic wiper (top corner) and lighting system. Viewed window (dashed red). Z direction (left) and X, Y direction (right). Dimensions in *mm*.

Due to restrictions in camera resolution at the respective speed, the cameras observe the active area of the filter medium over a window of 45×53 *mm* with a resolution of $86-90$ $\mu m/px$. In Subsection IV-A the confidence interval for depth values is found to be 94 μm. The stereo-optic method would be able to resolve in the range of the pixel resolution, provided that the distance of all points on the surface can be detected correctly.

III. Methods and Procedures

Once the setup is configured, images from both cameras can be captured. To obtain 3D information of the observed patch surfaces based on their distances to the filter, several post-processing steps are required. The filter will be removed after an experiment and may bend during an experiment but will maintain its position at the circumference once clamped in the holder and tightened. Therefore, calibration is required prior to an experiment. Here the applied methods are introduced in their experimental sequence.

A. Calibration

After the insertion of a clean filter medium sample, the camera system is calibrated by a planar calibration target with randomized dot pattern [10] [5]. Prior to filtration or cleaning one snapshot should contain the fixed filter with the pressing ring. The reason is given in Subsection III-C for the sake of operating sequence using the stereo-optical system.

B. Stereo-Matching

The goal of stereo-matching is to obtain a temporal sequence of depth maps in the stereo coordinate frame. The depthmaps need to be dense while conserving edges as good as possible. A specific challenge in the matching problem is the uncommon type of scenery, on which more recent learning-based stereo matching algorithms give worse results than traditional algorithms. Moreover, the pure white powder completely lacks texture, as does the background of the flying fragments, which is due to the directional illumination. For foreground fragments, this problem is mitigated by projecting a random dot pattern on the surface, but the background still remains dark and untextured. Therefore, any dense stereo matching algorithm will extrapolate the disparity information obtained at nearby fragment edges,thus reporting excessively high disparity in these regions. Adding a verification step through forward-backward matching helps to identify these regions, but nonetheless the fragment edges are not depicted sharply in the disparity map. After a qualitative comparison of recent stereo matching algorithms to traditional ones [3], we chose to apply a modified version of the well established semi-global matching algorithm (SGM) to compute the disparity map. A synchronized camera pair delivers images at a rate of 500 Hz. In contrast to the original SGM implementation, our algorithm is completely based on floating-point operations without performance loss and uses an M-Census metric for matching and bilateral filtering. To increase the overall accuracy, subpixel matching is performed over the entire image pyramid. Especially for planar, oblique objects like filter cake fragments, a slanted window matching is applied to compensate for perspective distortions. The obtained disparity information is converted to metric space with X and Y coordinates parallel to the primary camera's image plane and the Z coordinate towards the filter surface along the camera's principal direction. Figure 2 shows the result of two disparity map computations. A state of the art machine-learning-based method [9] (left) has a tendency

to hallucinate depths and does not give a clear distinction between individual patches. A modified SGM method with forward-backward-matching creates more sparse, but also more reliable 3D measurements.

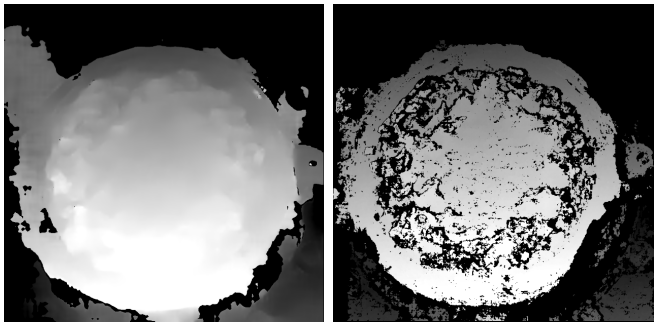


Fig. 2. Disparity map comparison of a machine-learning based stereo matching algorithm [9] (left) and a modified SGM algorithm (right).

C. Coordinate Transform Using Markers

The exact mounting position of the filter under test relative to the stereo setup may vary between experiments. In order to make these experiments comparable, a camera pose-independent coordinate system has to be defined and each experiment needs to be referenced to it. This reference coordinate system was defined relative to the mounting ring of the filter surface and is therefore as close as possible to the filter under test. Dedicated reference points with known world coordinates are marked on the fastening rings as shown in Figure 3, detected in the stereo images, refined to subpixel accuracy and finally triangulated. A rigid registration step transforms the measurement coordinate frame into the reference world frame. The resulting reference world frame is parallel to the filter plane with the perpendicular Z coordinate directing to the cameras and the origin at the center of the filter surface.

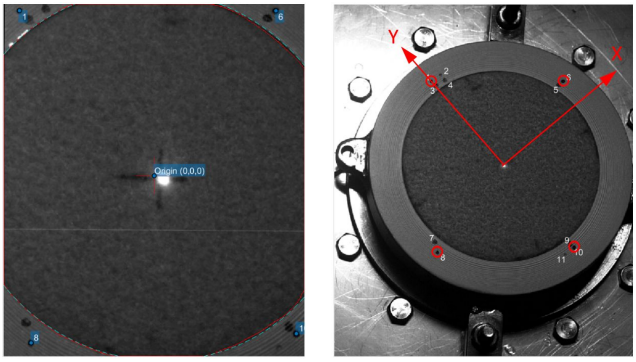


Fig. 3. Crosshair (left) and arrows (right) define origin and orientation of used camera pose-independent right-handed reference coordinate system. For calibration purposes, 11 dedicated reference points with known world coordinates on the filter fixing ring were defined (right). Red circles (right) respectively blue dots (left) mark the selected reference point subset during calibration routine – in particular case marker 1, 6, 8 and 10.

D. Annotation of Patches

The segmentation of individual filter cake patches is crucial to further assess their geometric parameters and temporal evolution (i.e. motion, further splits, collisions, etc.). A fully automated procedure is not deemed feasible, because it is a subjective measure when an individual fragment is completely detached from its neighbors, and how to assess partially occluded fragments. In addition, the visibility of patches may be affected by dust, and tilted stains may have a shaded surface that slowly fades into the black background. Therefore, AnnotationAssistant (an interactive segmentation software) has been designed for annotating patches semi-manually. The program takes a dataset of rectified images and let you select a specific image or step through the recorded image sequence. Fragments are annotated by marking a contour polygon. The process is assisted by an edge refinement between marked polygon points. Hereby, the connecting line between labeled polygon points is adapted to follow the local edges inbetween. Once annotated, filter cake fragments can be investigated based on automatically calculated geometric features. Figure 4 shows the first image where filter cake detachment initiates.

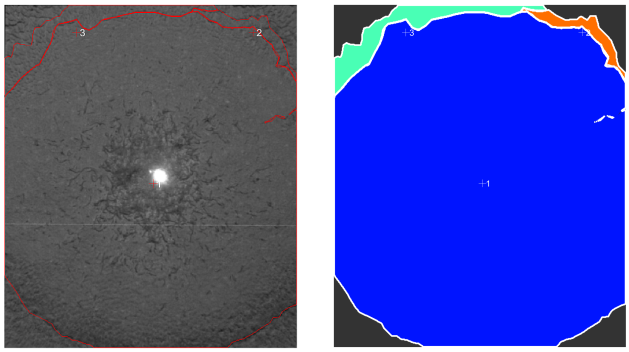


Fig. 4. First frame in recorded sequence showing the beginning of filter cake detachment. The left image shows annotated patches in this frame, gathered with the aid of AnnotationAssistant. The right image depicts the related label mask. Patch boundaries are schown in red.

Manual annotation is feasible for patches with length of at least 2 *mm*. Each patch automatically receives a unique identifier and is stored in a database. It can be refined and edited later on. For efficiency, labeled patches can be copied to the next time frame to be edited and adjusted to the new patch positions.

E. Extraction of Tracked Positions

In order to create consistent trajectories for segmented patches, they need to be tracked over time. Also here, automated approaches are deemed infeasible, as the patch motion between frames may be up to 20° in rotation and a quarter of the image area in translation. The appearance of a patch can change drastically between images, so local vicinity constraints cannot be applied. Rotation of patches in particular brings considerable appearance changes, so an explicit matching based on surface shape or contour is not feasible either.

The inroduced interactive AnnotationAssistant in Section III-D calculates geometric parameters and statistics, which are gathered per patch and allow quantitativ evaluation of the filter cake detatchment process. These parameters comprise patch center of gravity position, covered area, perimeter, orientation, several depth measures (real world z-position of a patch relative to filter fixation ring center) as well as the surface normal vector as a result of a plane fitting routine conducted on the 3D point cloud and some more statistics.

F. Distance from Filter Plane

The Z coordinates for an evaluated image can be seen in the corrected depth map on the right side of Figure 5. To obtain the course of detached patches, the image sequences are reduced to relevant ones that represent the detachment process. Patches may emerge, break into smaller patches, merge again or leave the observation window. The evolution of the patches can be represented, for example, in the form of a dendrogram, where the branch length represents the image sequence. Including the patch position, the paths of all patches can be ploted in a single, 3D space representing graph which requires renumbered patches [1].

IV. Experimental Results

In this chapter the quality of the measured geometric features of tracked patches is analyzed. First, the accuracy of the depth delivered by the application is determined. Then, limitations of reliability are indicated for certain conditions during cleaning. Finally, the propagation of the patches in Z direction is plotted and the degree of regeneration is calculated. Confidence intervals are given for 95% significance.

A. Depth Accuracy Considerations

Considering a surface resolution of $86-90\ \mu m/px$ and a depth to baseline ratio of $4:1$, a stereo matching accuracy of $\pm 0.5 px$ will translate to a depth error of ± 0.18 *mm*. Compared to the theoretical accuracy obtained by stereo-optic algorithms, the accuracy reachable in practice is lower, due to surface roughness and porous structure of filter fibers. The metallic surface of the ring reflects a good coordinate transform to the filter plane in the figures that follow. This plane may be tilted slightly, if the calibration points chosen at the markers differ in notch depth. For a fresh filter placed flat in the holder and fixed tightly, there is outward bulging as in Figure 5. The metal surface fluctuates in the range of -0.059 to 0.597 *mm*. Because of indented markers the true filter plane lies at a mean depth of 0.227 ± 0.041 *mm*. Near the center the fresh filter, i.e., a needle felt of untreated surface, fluctuates between 1.06 and 1.38 *mm*, a considerable surface roughness.

However, after filtration the deposited particulate layer differs radially in height as shown in Figure 6. The particulate surface on top of the metal ring is 0.19 to 0.785 *mm* away from the calibrated filter plane, which fluctuates in the same range as before, without particulate layer. The mean depth from the filter plane at the metal ring is 0.55 ± 0.047 *mm*. So the particulate layer adds 0.323 ± 0.064 *mm* thickness on top of the true filter plane. Layers on top of the metal ring may inherit the imperfections from the metal ring. On top of the filter far away from its circular edge and unintentionally created crater at center (due to large filtration velocity), the particulate surface fluctuates between 1.6 and 1.95 *mm* as a measure for surface roughness. Averaging over the regarding areas along the profile could smoothen the respective areas on top of the filter ranging from 1.63 to 1.85 *mm* in depth.

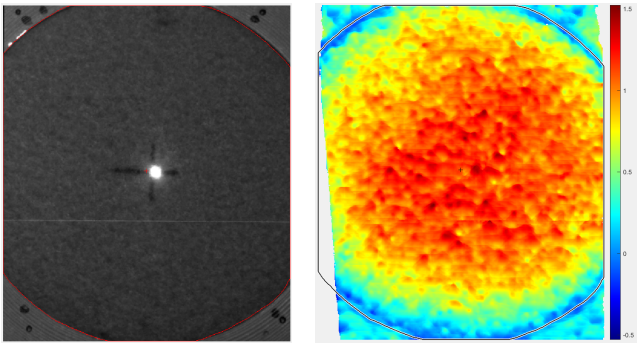


Fig. 5. Left: Image of a fresh, particle-free, flat clamped filter. The reference filter area is outlined red. Black dots represent reference markers. Right: Corrected auto-scaled depth map of the fresh filter area in *mm*. As can be seen, the filter bulges out.

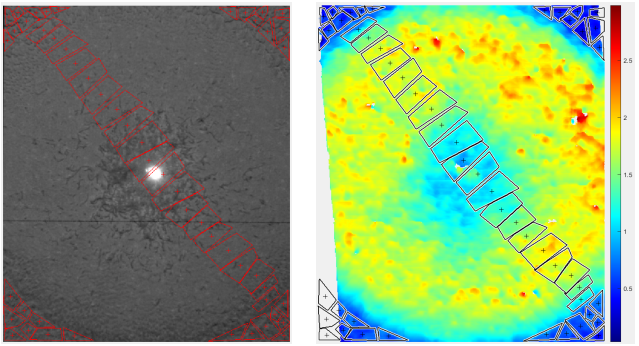


Fig. 6. Depth of particulate layer after filtration, in auto-scaled mm (right). Annotations for top of metal ring and radial profile (red outlines). See the fluctuations beside a radial distribution.

The flattest surface is on top of the metal ring; therefore, the accuracy of depth is 0.094 *mm* and may change for the thickness, here 0.128 *mm* (confidence width).

B. Limitations of Observation

Detachment of particulate matter is indicated in the first instance by breakups on (and in) the surface. Onsets of cracks on a surface are hard to detect, humans may assess from later crack propagation. Figure 7 shows the first breakup on the surface that later leads to patch detachment. More breakups occurred before a patch liberated from the filter surface. For comparison reasons the depth scale was manually set to the same range as in Figure 7 to Figure 10.

Liberated patches obey gravity but move to the top of the image as in Figure 7 due to test rig misplacement where the right side became the bottom side in this exemplary experiment. The cameras and so the images kept their relative

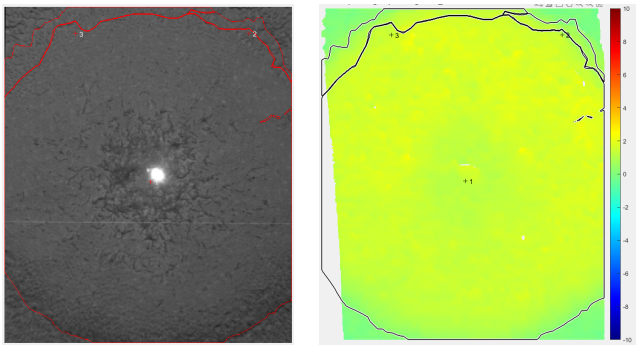


Fig. 7. Depth map at onset of detachment, in mm (right). Numbered patches (red outlines). Figure out surface cracks and see the point markings outside of concave patches 2 and 3.

position. The patches rotate and do not have sharp edges, a human may regard areas subjectively as patches or not, as is shown. Patches overlay each other and tilt at the surface. Sharp jumps in distance cause stereo-matching to fail. For instance, a projection of a distant object is clear at one viewing angle but the projection of the same object from a different viewing angle hides a spot seen with the former viewing angle where stereo-matching fails. It may wrongly produce a counter-copy for large distant objects of inverse increased distance like the blue one at the bottom right half.

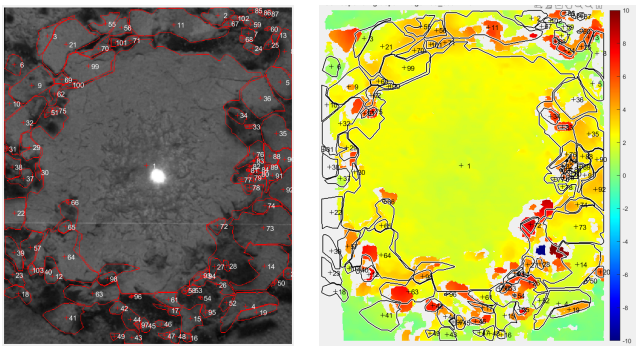


Fig. 8. Annotated and numberated patches at end of liberation process (left) and corresponding depth map based on mm (right).

The mismatch may be clearer from an example of the two viewing angles. At the left half of Figure 9 there may be an object within the blue rectangle. At the right half the scene is viewed from an angle at the right and the object captured within the blue rectangle. The position of the object differs in both images, where the depth at the image position of the back camera seems to be correct and the counter-copy is situated at the image position of the front camera. A small, resolved particle causes the stereo-matching to fail as well, highlighted are areas with numerous closely resolved particles. Smaller, unresolved particles may smear within a pixel to altered brightness and color.

Patch 6 at the top left is surrounded by particles so small (1-3 pixels) that one may interpret that as fog of dust. At this intermediate condition stereo-matching fails at some locations of the dust. If the objects are smaller, they blur the

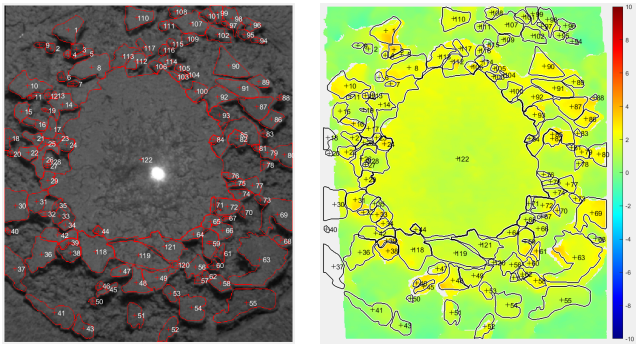


Fig. 9. Depth map at the end of cleaning, in mm (right). From camera view, gravity goes up. Numbered patches (red outlines).

background and stereo-matching generates the background depth, not shown here. An event of a dense haze of dust was not found, therefore it is not clear so far whether stereo-matching can cope with haze. Matched surfaces are subject to little random fluctuations among an image sequence.

C. Degree of Regeneration Analysis

At the end of cleaning, patches have either completely detached from the filter surface, have remained at their place or redeposited at the surface. An example is shown in Figure 10. In a conservative (safe assumption) view of non-regenerated area, tilted patches are considered by their unprojected surface area and areas of covering patches are considered multiple times. Those small areas were neglected where a normal vector was not provided.

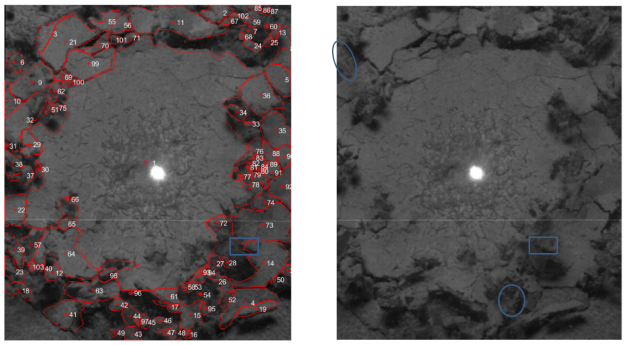


Fig. 10. Stereo images of back (left) and front (right) camera at end of detatchment process. From camera view, gravity goes up. All annotated patches are numbered (red outlines, left). Blue rectangles denote examples of tilted patches that lead to false area results and blue ellipses mark examples of small particles.

The degree of regeneration is the ratio of liberated quantity to initial quantity, here quantified by area. Using projected areas alone gives 39.8% [1]. The conservative approach gives 20.2%.

D. Patch Trajectory Analysis

To plot the distance that the patches travelled the individual patches are abstracted in Figure 11 by their projected radial position from the filter center. The individual patch areas are indicated by the size of the circles. To ascertain a

trend a few consecutive images were selected. During the cleaning air pulse, the filter moves from the initial stage (image 000000) to the final stage (image 004002), where the radial profile of the surface was obtained from the deposited patches. This reflects the crater of almost no dust at the center and the bulge caused by the pulse of pressurized air. Strictly speaking, the measurements are from the top of the patches. Compared with Figure 10 the circumference at about 20*mm* is mostly of tilted patches with larger surface dimensions than thickness. Therefore, the radial profile fit to their mean depths overestimates the filter surface more than the fit to mean depths of plane patches. The patch at the center is the largest occupying a radius of 11*mm* the least as other patches appear at larger radii. Some of them may have redeposited on top of the central patch which requires further investigation. Prior to image 000219 the large patches have existed since the detachment. Over the course of a patch movement, it may appear in the selected images for a short while and disappear from the plot. As filtration was discontinued during cleaning, the patches experience only the force exerted by the compressed air on the filter and follow that path afterwards (up arrow). For this experiment the test rig was placed with the right side at the bottom, therefore the filter is oriented vertically and gravity acts on the patches laterally to the filter plane. Thus, the particles follow a rather parabolic path in the plot. The pulse lasted longer than the images were captured. However, one patch seems to redeposit (down arrow), outside the filter.

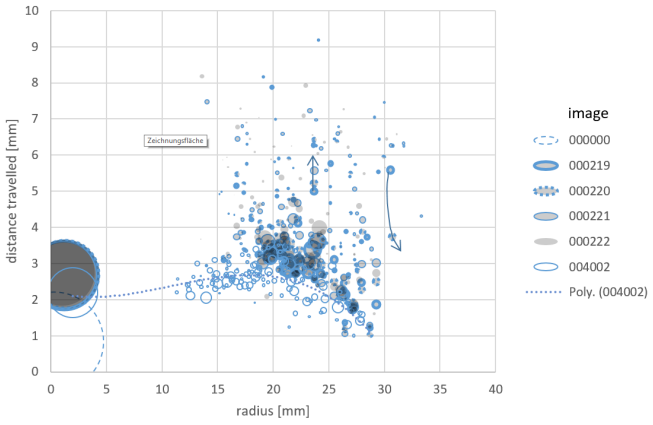


Fig. 11. Travelled distance between 22^{nd} and $28^{th}ms$ of detachment (000219 to 000222, grey of various outline) against radius from filter center, plus initial (000000 of dashed outline) and final stage (004002 of solid outline). Circle size indicates patch area. Final surface (dotted trendline); at final stage all shown patches are deposited on filter.

V. Conclusions

In total, 77 experiments have been conducted, with the filter cleaning process captured on a series of 15 to 20 images per experiment. The presented computer vision setup and analysis system allows for the first time a quantitative, three-dimensional analysis of dust filter cleaning processes at 500 Hz. It is evident, that errors in size and orientation of patches can occur on an individual level due to mismatches and due to the patch shape. Typically, patches have an area of 5 $mm \times 5$ mm, with a thickness smaller than 2 mm. If a patch rotates and is viewed from the side, its size will be grossly underestimated. The last Figure 10 shows that it is possible to distinguish between redeposited patches and patches that remain at the filter at a size-based level.

VI. Outlook

As a result of this work, the cumulative area distributions can be obtained from the experiments and will be compared with data from literature. Also, the validation of simulations will become possible in terms of motion paths of the patches, and velocities of the patches at a given time, as soon as above measures allow error-tolerant tracking of individual patches. Based on this data set, the next step is to create evolution-trees of patches, i.e. to track the decomposition of patches into smaller ones and establish relationships to ancestors and predecessors. To identify the class of patches that travel long distances, more cumulative analyses will assist, in terms of their size and local origin. The labeled dataset of fragments will act as a reference in the future to train an automated instance segmentation, where the now laborious patch labeling will be replaced by a per-pixel classification step. In occluded regions, where the matching of two cameras failed, additional cameras may assist, i.e., a multi-camera-view, especially in the matching of layered objects. Besides the tracking of patches, by the above measures, the local and time-resolved investigation of powdery layer breakage may become possible. Finally, the movement of flexible filters or the compaction of powder layer may be determined separately.

References

[1] V. D'Ercole, "Study of movement and dimension of filter cake patches during pulse jet cleaning by high-speed stereo camera for identifying optimal cleaning conditions," Master's thesis, Graz University of Technology, 2021.

[2] A. Dittler, B. Gutmann, R. Lichtenberger, H. Weber, and G. Kasper, "Optical in situ measurement of dust cake thickness distributions on rigid filter media for gas cleaning," *Powder Technology*, no. 99, pp. pp. 177–184, 1998.

[3] H. Hirschmüller, "Semi-global matching – motivation, developments and applications," *Photogrammetric Week*, 2011.

[4] C. Kanaoka and M. Amornkitbamrung, "Effect of filter permeability on the release of captured dust from a rigid ceramic filter surface," *Powder Technology*, vol. 118, pp. 113–120, 2001.

[5] M. Klopschitz, G. Lodron, G. Paar, and B. Huber, "Line processes for highly accurate geometric camera calibration," in *Proc. AAPR*, 2017.

[6] F. Löffler, H. Dietrich, and W. Flatt, *Dust Collection with Bag Filters and Envelope Filters*, H. Simon, Ed. Springer Fachmedien Wiesbaden, 1988.

[7] M. Rüther, M. Saleem, H. Bischof, and G. Krammer, "In-situ measurement of dust deposition on bag filters using stereo vision and non-rigid registration," *Assembly Automation*, vol. 25, no. 3, pp. 196–203, 2005.

[8] M. Rüther, M. Uray, H. Bischof, G. Krammer, and M. Saleem, "An optical measurement device for evaluating dust deposition on flexible filter surfaces," in *10th Computer Vision Winter Workshop CVWW 2005*, 2005.

[9] G. Yang, J. Manela, M. Happold, and D. Ramanan, "Hierarchical deep stereo matching on high-resolution images," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[10] Z. Zhang, "A flexible new technique for camera calibration," *IEEE transaction on pattern analysis and machine intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.

Enabling Classification of Heavily-occluded Objects through Class-agnostic Image Manipulation

Benjamin Gallauner[1], Stefan Thalhammer[2] and Markus Vincze[2]

***Abstract*— Image classification is a fundamental task of computer vision. When training classifiers on images of heavily-occluded objects, classification is strongly influence by the appearance of the occluders. That leads to a severe drop in classification accuracy when confronted with unknown occluders. More precisely, when classifying shelf types in a shop floor, occluded by household items, the full range of diversity of those occluders has to be regarded as unknown for test time. However, resulting in a severe drop in classification performance when dealing with images containing unseen occluders during training time. In order to improve classification, we exploit the generalization capability of unknown object instance segmentation. We segment and replace the object appearance of the unknown occluders with random intensity noise. Consequently, the classifier is able to focus on those image parts containing the objects of interest. We show the theoretical foundation of our approach through empirical analysis on a test set with large data distribution shift with respect to the training set.**

I. INTRODUCTION

Image classification is a long standing challenge in computer vision. It refers to the task of assigning one or a distribution of classes to a given image [11]. Classification, as fundamental computer vision task, is often used to benchmark network architectures and domain adaptation. In computer vision systems classifiers can help to provide priors for subsequent stages.

This paper is concerned with the special case of classifying objects that are heavily occluded. In particular, we aim to classify images of shelves belonging to one out of three classes (standing, hanging and bucket). The respective shelves are part of a shop floor and thus heavily-occluded by a broad variety of household objects. Figure 1 shows a representative sample of the class *Bucket* and an overview of our proposed approach to solve the problem at hand. Training a classifier on the available images induces a bias such that class predictions are primarily made by memorizing the occluding objects. In order to guide prediction making towards leveraging image information belonging to the actual descriptive parts of the image, i.e., the shelves, the occluding object information has to be removed. Since the occluders are considered to be unknown during test time, those have to be treated as unknown.

This work has been supported by the Austrian Research Promotion Agency in the program ICT of the Future funded project Knowledge4Retail (FFG No. 879878) and the Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation, and Technology (BMK)

[1]Benjamin Gallauner is with the TU Wien, 1040 Vienna, Austria e01631744@tuwien.ac.at

[2]Stefan Thalhammer and Markus Vincze are with the Faculty of Electrical Engineering and Information Technology, TU Wien, 1040 Vienna, Austria {thalhammer, vincze}@acin.tuwien.ac.at

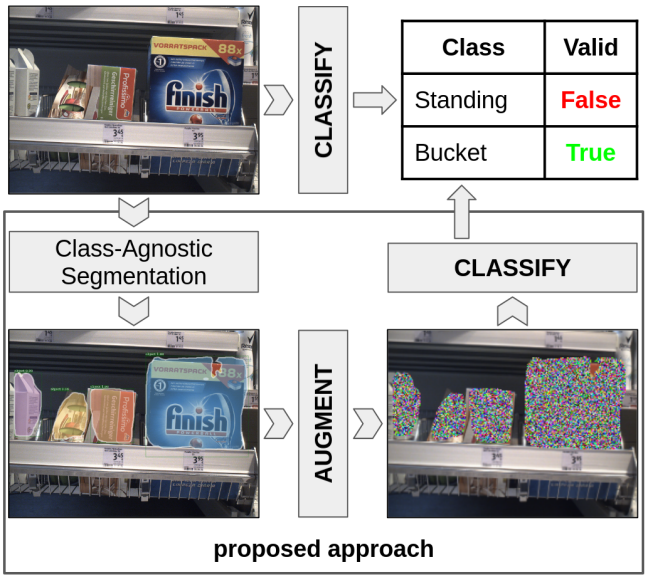


Fig. 1. Approach overview: Via class-agnostic segmentation we guide class prediction towards focusing on relevant image parts.

Existing approaches try to learn a general concept about objects, for recognizing unseen and unknown objects [8], [15], [16], [17]. These approaches extract different object information and employ different strategies based on the given problem. We are interested in removing the image information of occluding objects, thus we require instance segmentation. While [15] and [16] provide segmentation masks for unseen objects, these methods require depth and RGB-data for very constraint scene setups: Segmentation is provided for objects on a table plane from a top view. This method is not guaranteed to generalize to objects in arbitrary placements and varying backgrounds. The authors of [8] provide a method for detecting objects from RGB images based on learning the general concept of "objects" but does not provide instance segmentation. In [17], instance segmentation is provided for a broader range of objects but a few images of the objects to segment are required for fine-tuning their approach.

In this work, we are interested to learn the general concept of "object" in a way that it generalizes to unseen objects without requiring images of the involved objects, while also performing instance segmentation. As such, we learn to produce a joint encoding of diverse household objects, using objects belonging to the categories household, kitchen, tool and shape from the YCB dataset [2]. This is done by

assigning the same class to all objects involved, effectively learning to separate object instances from background. Using that encoding, we are able to eliminate the foreground information that mainly consists of household objects to improve classification performance of our heavily-occluded objects of interest.

The remainder of the paper discusses related work in Section II, provides the problem statement in Section III, which is followed by a description of our proposed approach in Section IV and evaluations in Section V. Lastly, Section VI concludes the paper.

II. Related Work

Image classification is a fundamental problem in computer vision [9]. As such it is often used to benchmark feature learning [5], [10], [12] and domain adaptation [1], [3]. In this paper we are interested in solving the challenging problem of classifying objects that are containers of smaller objects and thus heavily occluded by these. If the background is less dominant than the foreground, foreground-background separation or salient object detection approaches can be used for the separation [13]. However, the classification of images where the majority of the image is comprised more by occluders than by the object of interest is highly challenging when no respective annotations are available. In order to distinguish the occluders from the objects of interest, detection or segmentation of unseen objects can be used to synthesize the required annotations.

Class-agnostic object detection is meant to draw bounding boxes around image regions containing potential unknown objects [8]. If a few images of the occluding objects are annotated with masks, few-shot instance segmentation learning approaches can be applied [17]. Alternatively, the availability of annotated data from similar object instances of the same category enables the learning of the ability to segment unseen objects [15], [16]. This, however, requires the knowledge of which object categories are to be expected in the test images.

We aim to generalize instance segmentation to a a broader range of objects. Thus, we combine class-agnostic object detection and unseen object segmentation in order to achieve instance segmentation coming from multiple categories of objects.

III. Problem Description

Training classification for objects that are heavily occluded leads to learning to solve the task using all of the image. This, in turns, leads to feature extraction focusing on extracting any set of features that minimizes the task loss for the given training set. However, there is no guarantee that the extracted features and the decision function yields generality. Table I presents the recall for successfully classified images on a set of images of shelves from the same shop floor split into *Training* and *Val*, and one *Test* set captured in a different location. The *Test* set features different occluders that are unseen during training time. More information on the image sets is provided in Section V-B.1.

TABLE I
Classification Recall on *Val* and *Test*.

Set	bucket	hanging	standing	average
Val	1.0	1.0	1.0	1.0
Test	0.08	1.00	0.09	0.39

A significant drop in performance is observed from *Val* to *Test*. We want to emphasis that since we have a 3-class problem, the classification recall on *Test* is close to random output. Thus, the network learns no generalized encoding relevant for the problem to be solved.

IV. Approach

Given a set of images X, each featuring a class $C \in \{1,...,n\}$ of interest, where the image parts describing the class information are largely occluded by a set of unknown objects U. We employ a function $\hat{y} = f(x)$ to provide segmentation masks y for U in x. Subsequently, x is augmented with $x\{\hat{y}, p\} = \mathscr{U}\{0,...,255\}$ to eliminate the object information of U in x. Where p are the pixels in the mask and $\mathscr{U}$ is a uniform distribution. In order to learn a function $\hat{c} = g(x)$ for image classification.

A. Unknown Object Instance Segmentation

In order to generalize instance segmentation to arbitrary objects O, the input data x has to be composed of an object set $\hat{O}$ sufficiently large and sufficiently diverse to encode a feature space that effectively interpolates between object instance $o_{1...n}$. Thus, $\hat{O}$ has to be chosen in such a way as to provide a superset of U. Since it is intractable to provide the whole variety of object types in x, we choose $\hat{O}$ to provide samples of all the expected object categories, in order to learn interpolation between objects. We learn a function $\hat{y} = f(x)$, where $\hat{y}$ are the masks of the object instances $o_{1...n}$. Class-agnostic segmentation $f(x)$ is enabled by providing $\forall o \in \hat{O} : o_n = o$. In other words, we map all of $\hat{O}$ to the same object class. Thus, learning $f(x)$ to separate foreground objects from background and segmenting the foreground by finding instances.

B. Classifying Heavily-occluded Objects

To facilitate classification of heavily-occluded images, we apply f to X. The resulting $\hat{y}$ provides segmentation masks for U. The instance segmentations y are subsequently used to augment training images so that $x\{\hat{y}, p\} = \mathscr{U}\{0,...,255\} : \forall o \cap \forall x \in X$, thus, replacing foreground object information of U with random pixel intensities. The resulting augmented image set X_a is used to learn the function $\hat{c} = g(x)$. By eliminating the object appearance of U from X, g can focus on encoding features relevant for predicting c given x_a.

V. Experiments

The following section provides implementation details and experiments. Quantification of the functioning of our approach is done by providing comparison to standard techniques for improving image classification performance.

A. Class Agnostic Segmentation

To facilitate class agnostic object instance segmentation of a broader category of objects, $\hat{O}$ has to be chosen to represent the variations of the expected objects in the test set.

1) Segmentation Data: Since $f(x)$ is expected to encode the concept of an "object", the training data x has to be chosen that the corresponding y is given in a way that a clear distinction between foreground objects and background exists. The expected unknown occluding objects are household items. As such, training data for f has to be chosen to reflect the diversity of object appearances with U being household items. Care has to be taken that the variations in x with respect to aspects such as object placement and interaction, as well as illumination and contrast are sufficient to generalize to the domain of X. The YCB-video dataset [14] features 21 objects derived from the YCB-dataset [2]. The objects in YCB-video belong to the categories food, kitchen, tool and shape items with diverse setup and scene illumination, thus, representing diverse object appearances. YCB-video consists of 92 videos containing 133,827 frames. These are split into 113,199 training and 20,628 validation images.

2) Segmentation Training: In order to show-case the generality of our class-agnostic segmentation approach we fine-tune the standard approach for instance segmentation, Mask-RCNN [4] with Resnet101-backbone [6] pretrained on ImageNet [11], for encoding $f(x)$. As such, showing that no specialized network configuration is required, to generalize to unknown objects. Training is done for one epoch with a base learning rate of 0.001. The loss is reduced by one magnitude after 66% and 90% of training iterations, which correlates with the standard schedule. All 21 YCB-video classes are trained to be the same class. Consequently, Mask-RCNN has to learn the common traits that describes an object based on the YCB-video objects. As a result, we train to predict anchor locations containing an object of interest, while simultaneously predicting per-pixel instance segmentation for each positive anchor. Non-maximum suppression is applied to circumvent multiple detections of the same object.

3) Class-Agnostic Segmentation Results: Figure 2 presents exemplary class-agnostic segmentation results for our shelves training set and YCB-video. On YCB-video, the results indicate that a joint latent representation is encoded by $f(x)$. The mean Average Precision (mAP) for Intersection-over-Union-thresholds (IoU), from 0.5 to 0.95 with a step size of 0.05, is 0.714 for object detection and 0.676 for object instance segmentation. Instance segmentation is also predicted on unseen images of the involved objects. The middle row in the right column also shows a properly segmented background object that is not annotated in the training set. An error case occurring on the images of shelves is visible in bottom image of the right column. Showing a segmentation mask that includes the edge of the table connected to the tuna can standing on it. Similar errors are observable in the shelves images in the left column showing objects not contained in YCB-video. For these, the price labels attached to the shelves are often detected as separate objects or via segmentation masks

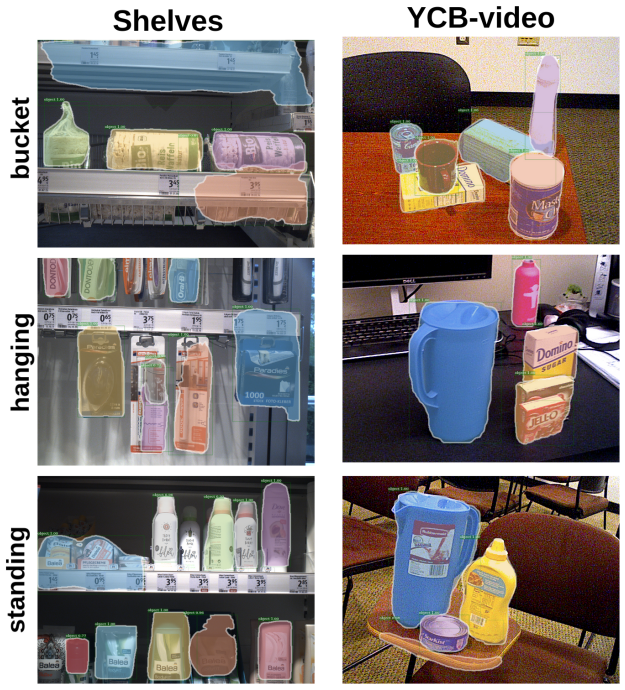


Fig. 2. Class-agnostic instance segmentation results on images of our training set for shelves (left column) and on unseen images of YCB-video [14] (right column).

Fig. 3. Example images of the sets *Train* and *Val*, and *Test*.

connected to one. This behavior is acceptable since price tags do not provide useful cues to distinguish between shelf types. Unknown objects are segmented in most of the cases. As such, providing a useful basis for eliminating foreground information from the images to train $g(x_a)$ on. Enabling $g(x_a)$ to focus more on background information of x.

B. Image Classification with Heavy Occlusion

Having an $f(x)$ for providing $\hat{y}$ of U image manipulation can be applied to X in order to create x_a.

1) Classification Data: Training and validation data is collected in the replica of a shop floor with limited variation regarding occluders. The procedure is automated such that a camera is mounted to a robotic arm that pans down in front of the cupboard containing shelves. Since the aim of this work is to generalize to a broad variety of potential scenes, a *Test*-set is captured in an actual shop floor accessible to

customers. The shelves have varying characteristics in the designated sets:

- *Hanging* provides distinct-shaped hooks to hang product. These have the same shape and color in all sets.
- *Standing* provides storage spaces for different product separated with transparent plastic dividers. Those dividers have the same color in all sets, but the shape is slightly different, exhibiting a straight edge in *Train* and *Val*, and a chamfered edge in *Test*.
- *Bucket* provides bin-like storage cages with severe differences in shape and color. Those in *Train* and *Val* have a metallic appearance, inclined front and price tags attached to the buckets, while those in *Test* are matt white, exhibiting a straight front and the price tags are attached beneath the buckets.

Exemplary images of the sets are presented in Figure 3. Set sizes are 486, 122 and 106 images for *Train*, *Val* and *Test*, respectively.

2) Learning Classification: For classification, Resnet50V2 [5] without pretraining is used as $g(x)$. Training is done for 50 epochs on *Test* on the 3-class problem. We use a learning rate of 10^{-4}, optimizing with stochastic gradient descent and cross entropy loss.

3) Classification Results: The standard approach to generalize to novel domains is to augment the training data. Applying geometric and color space augmentations virtually increases the training data and decouples estimation making from some characteristics of the training set.

Thus, we define applying image augmentations to X as a baseline. In Table II an ablation regarding our applied augmentations an their influence on the classification performance is reported using the classification recall. In order to guide the network towards more effective discrimination of image classes we apply MixUp-Augmentation to our training data. Results are provided for the test set.

TABLE II

ABLATION WITH RESPECT TO AUGMENTATIONS AND THEIR RESPECTIVE INFLUENCE ON THE CLASSIFICATION RECALL.

Augmentation	bucket	hanging	standing	average
None	0.08	1.0	0.09	0.39
zoom (20%)	0.17	0.77	0.43	0.46
rotation (15°)	0.20	0.72	0.42	0.45
translation(10%)	0.23	0.75	0.32	0.43
shear($\lambda = 0.1$)	0.12	0.96	0.29	0.46
horizontal flip	0.10	0.99	0.28	0.46
brightness(10%)	0.25	0.79	0.29	0.44
MixUp [7] ($\alpha = 0.2$)	0.26	0.85	0.34	0.48
all	0.42	0.95	0.12	0.50

For unknown object instance segmentation we have to set a detection threshold for $f(x)$. Table III compares different detection thresholds using grid search. The intuitive and usually generally applicable value of 0.5 provides the best results in terms of average recall over all 3 classes.

Table IV provides results comparing our approach using a detection threshold of 0.5 to standard augmentations. The last two columns provide the average over all three classes (2nd to the right) and classes *Hanging* and *Standing* (rightmost). For our approach we use a detection threshold of 0.5 for segmenting and augmenting unknown objects. Averaged over all three classes standard augmentations result in a higher recall than using our manipulated training data x_a. Considering the classes with little to no difference in appearance in *Train/Val* and *Test*, *Hanging* and *Standing*, our approach significantly improves over standard augmentations. Our approach does not classify the *Bucket* of *Test* as such. Which is to be expected due to the severe difference in appearance between *Train/Val* and *Test*. This behavior hints that the network is able to focus more on the relevant background data and spatial relations of the scene, while focusing less on occluding objects. Combining standard augmentations and our image manipulation bridges the performance gap between using only standard augmentations and our image manipulation.

TABLE III

COMPARISON OF DIFFERENT DETECTION THRESHOLDS FOR SEGMENTING AND MANIPULATING TRAINING DATA, EVALUATED ON THE *Test*-SET USING THE AVERAGE RECALL.

threshold	bucket	hanging	standing	avg
0.3	0.03	0.64	0.51	0.39
0.4	0.01	0.60	0.61	0.41
0.5	0.02	0.65	0.58	0.42
0.6	0.00	0.59	0.63	0.41
0.7	0.00	0.63	0.51	0.38

TABLE IV

COMPARISON OF DIFFERENT STRATEGIES FOR TRAINING DATA MANIPULATION FOR SHELF CLASSIFICATION PRESENTED AS CLASSIFICATION RECALL.

Aug.	bucket	hanging	standing	avg(all)	avg(2&3)
None	0.08	1.0	0.09	0.39	0.55
all aug.	0.42	0.95	0.12	**0.50**	0.54
ours	0.02	0.65	0.58	0.42	**0.63**
ours+aug.	0.04	0.26	0.95	0.46	0.61

VI. CONCLUSIONS

We present an approach for removing unknown objects from images to improve the classification performance on the objects of interest that are occluded by the unknowns. Further investigations will investigate how significantly class-agnostic segmentation can improve classification performance on highly occluded objects. As such, test data with more various and diverse object sets as training and test data for class-agnostic segmentation and classification will provide useful insight. The research performed in this work focuses on household objects. We aim to extent the proposed approach to arbitrary unknown object instance segmentation to facilitate broader applicability in more diverse domains. As such, promising future contributions could be made to open set recognition and learning new objects online.

ACKNOWLEDGMENT

We acknowledge the contribution of Vanessa Hassouna, Alina Hawkin and Simon Stelter who are with the Institute for Artificial Intelligence, University of Bremen, for capturing data in their shop floor and acknowledge DM Drogerie Markt GmbH for providing items and shelves.

REFERENCES

[1] S. Ben-David, J. Blitzer, K. Crammer, F. Pereira, *et al.*, "Analysis of representations for domain adaptation," *Advances in neural information processing systems*, vol. 19, p. 137, 2007.

[2] B. Calli, A. Singh, J. Bruce, A. Walsman, K. Konolige, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Yale-cmu-berkeley dataset for robotic manipulation research," *The International Journal of Robotics Research*, vol. 36, no. 3, pp. 261–268, 2017.

[3] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[4] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.

[6] ——, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[7] Y. N. D. D. L.-P. Hongyi Zhang, Moustapha Cisse, "mixup: Beyond empirical risk minimization," *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=r1Ddp1-Rb

[8] A. Jaiswal, Y. Wu, P. Natarajan, and P. Natarajan, "Class-agnostic object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 919–928.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[10] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," in *NeurIPS*, 2019.

[11] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[12] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[13] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: An in-depth survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[14] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," 2018.

[15] C. Xie, Y. Xiang, A. Mousavian, and D. Fox, "The best of both modes: Separately leveraging rgb and depth for unseen object instance segmentation," in *Conference on robot learning*. PMLR, 2020, pp. 1369–1378.

[16] ——, "Unseen object instance segmentation for robotic environments," *IEEE Transactions on Robotics (T-RO)*, 2021.

[17] C. Zhang, G. Lin, F. Liu, R. Yao, and C. Shen, "Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5217–5226.

Real Estate Attribute Prediction from Multiple Visual Modalities with Missing Data

Eric Stumpe[1], Miroslav Despotovic[2], Zedong Zhang[2] and Matthias Zeppelzauer[1]

***Abstract*— The assessment and valuation of real estate requires large datasets with real estate information. Unfortunately, real estate databases are usually sparse in practice, i.e., not for each property every important attribute is available. In this paper, we study the potential of predicting high-level real estate attributes from visual data, specifically from two visual modalities, namely indoor (interior) and outdoor (facade) photos. We design three models using different multimodal fusion strategies and evaluate them for three different use cases. Thereby, a particular challenge is to handle missing modalities. We evaluate different fusion strategies, present baselines for the different prediction tasks, and find that enriching the training data with additional incomplete samples can lead to an improvement in prediction accuracy. Furthermore, the fusion of information from indoor and outdoor photos results in a performance boost of up to 5% in Macro F1-score.**

I. INTRODUCTION

Over the last few years, significant progress has been made in the field of automatic real estate appraisal. While earlier models have exclusively utilized textual and categorical input data such as the number of rooms or the floor area [4], [20], [28] to predict building attributes, recent research has demonstrated that the inclusion of visual information from building photographs can be beneficial [21], [14], [29]. Examples include sophisticated price estimation models [21], machine learning methods for predicting building heating energy demand [7], but also the analysis methods for architectural style [8]. A prerequisite for the development of efficient machine learning models in the domain of automatic real estate valuation is the availability of a sufficiently large and well-annotated dataset. In practice, obtaining enough data is usually not an issue, but the corresponding annotations are often incomplete or include varying annotation categories/schemes when obtained from different sources. This calls for new automated methods to fill such annotation gaps and missing data.

In this work[3], we leverage the information contained in real estate images to predict high-level real estate attributes and thereby show a novel way to fill missing data in real estate databases. Examples for such attributes that we examine are e.g. the type of commercial use of an object (e.g. "industrial", "hospitality", "retail" or "office") or the general type of a building, i.e., whether it is a commercial building or a residential building. Specifically, we use pairs of facade and interior photos of real estate objects as input which we refer to as two different visual input modalities in the following. This means that the input to our method is a pair of indoor and outdoor images, see also Figure 1. The facade and interior embody separate visual aspects of the same property and contain complementary clues for estimating a particular attribute. Consider the photo pair of Figure 1 as an example for the task of differentiating between commercial and residential real estate objects. The large window fronts of the facade image serve as an indicator that this object may be a commercial office building. Even stronger hints are provided by the many office chairs in the interior image. This example illustrates that for each of the two visual input modalities, different types of information need to be extracted and fused to successfully predict a particular attribute. To evaluate how this can be best achieved, in this work we implement and evaluate three multimodal architectures representing different fusion approaches with different fusion levels. In addition, interior and facade photos are not always both available for each real estate object. We therefore analyze how robust our proposed models are to missing modalities and whether using additional incomplete samples in the training set can improve prediction accuracy.

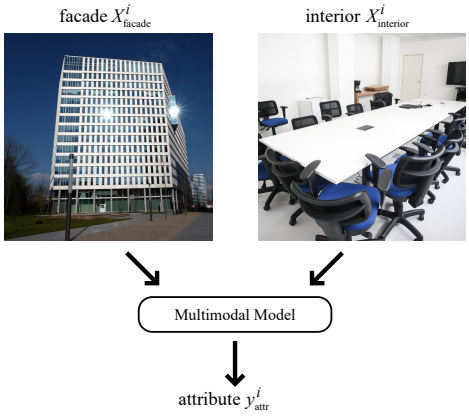


Fig. 1. Concept of multimodal learning with two visual modalities.

II. RELATED WORK

In this section we first provide an overview of computer vision methods for real estate analysis and then review re-

[1]E. Stumpe and M. Zeppelzauer are with the ICMT Institute of Creative Media Technologies, St. Pölten University of Applied Sciences, St. Pölten 3100, Lower Austria, Austria (estumpe@fhstp.ac.at; matthias.zeppelzauer@fhstp.ac.at)

[2]M. Despotovic and Z. Zhang are with the Kufstein University of Applied Sciences, Kufstein 6330, Tirol, Austria (miroslav.despotovic@fh-kufstein.ac.at; zedong.zhang@fh-kufstein.ac.at)

[3]This research was funded by the Austrian Research Promotion Agency (FFG) project 880546 "IMREA" and we are grateful to DataScience Service GmbH for providing the data.

lated work on multimodal image classification and prediction from missing data/modalities.

A. Real Estate Image Analysis

An early approach on multimodal learning for real estate analysis, which also utilizes visual information, was proposed by Ahmed et al. [1]. To leverage the image information of a building, the authors extracted SURF features [2] from different room types and trained a neural network to predict the price from both visual and textual features. In another work by Kostic et al. [14], image entropy, level of greenness, and features extracted from a CNN pretrained on ImageNet [6] were used for price prediction. A method for estimating the age of a building from its visual appearance was introduced by Zeppelzauer et al. [30] where the authors extracted patches of interest via SIFT features [17] and gave them as input to a neural network that predicts the building age through decision fusion. This method was extended in Despotovic et al. [7] for predicting the heating demand of a building. A model based on long-short-term-memory (LSTM) networks was developed by You et al. [29]. To achieve a robust estimate of a property's value, the LSTM network was also provided with photos from the neighborhood of the building. Bin et al. [3] took advantage of attention modules [26] and fused information from both textual data and satellite images in order to automatically predict property prices in Los Angeles. Using Crowdsourcing, Poursaeed et al. [21] built a dataset with luxury scores for different room types. Subsequently, a CNN network was trained to predict the luxury score of each room and merge it with textual data to predict the property price. A comprehensive overview of the emerging trend of image analysis in the real estate domain has recently been provided by Koch et al. [13].

B. Multimodal Learning

An important architectural design choice in multimodal learning is where to fuse the information from different input modalities. *Early fusion* models combine all modalities at the input level, which can be achieved by concatenating raw data or preprocessed input features [15], [11]. Limitations for this type of models can arise from differing dimensionalities and sampling rates of the input modalities [22]. Another option is to fuse modalities at the decision level of the model [10], [16], [19], which is usually called *late fusion*. In this case, a separate classifier is used for each modality, and the overall model prediction can be computed by using e.g. the maximum or average of the predictions or by stacking a meta-classifier on top. When the information of modalities is merged throughout the model, it is referred to as *intermediate fusion*. This type of fusion can be achieved in a variety of ways. Wang et al. [27] proposed a strategy for handling pairs of corresponding RGB images and depth maps. Based on the batch normalization activation levels of the model's intermediate layers, feature map channels are exchanged between both modalities to replace irrelevant information. The work of Nagrani et al. [18] has shown that Visual Transformers [9] can be successfully applied to a multimodal problem. To exchange cross-modal information in the model they used attention bottlenecks. In our study we apply the ideas of Joze et al. [12] for one of our three network variants. The authors used so-called multi modal transfer modules (MMTM) between modality-specific CNN streams. These modules help to recalibrate the magnitude of channel-wise features in each stream, which will be described in more detail in section III.

C. Missing Modalities

Sun et al. [23] proposed an image translation method that can compensate for the absence of single modalities. They implemented an encoder-decoder architecture for each modality and arranged them in a cyclical structure during training so that one image modality can always be reconstructed from the encoded information of another modality. In a similar approach, Tran et al. [25] developed a cascading network of residual autoencoders for the task of predicting missing modalities. Choi et al. [5] used subnetworks for each modality, each yielding a feature vector of the same dimension. Then, a random sampling process is applied which takes sparse features from each modality and combines them, improving the ability of the network to compensate for missing information. In our work, the ability of our models to handle missing data is not achieved through the network architecture design, but through data augmentation.

III. Approach

The main goal of our work is to develop a network architecture that can perform the following functions.

1) When provided with an input pair of both a photo of the building facade X^i_{facade} and from the interior X^i_{interior} of the same real estate object i, it should be able to predict the correct class y^i_{attr} of a given category (see Figure 1).
2) The model should be capable of dealing with missing modalities, which in this instance refers to either an absent indoor X^i_{interior} or facade photo X^i_{facade}.

In our method, we handle a missing modality by representing the missing X^i_{interior} or X^i_{facade} as a black image with all RGB values set to zero. We further investigate how different fusion strategies perform in this scenario. To this end, we implement three model architectures, each representing a different fusion archetype. A full description of these architectures can be found in Section III-A. The high level attributes which we investigate are the commercial type, residential type and object type of a property. More details on these attributes can be found in IV-A To evaluate our approach, we formulate the following five research questions (RQs), which we will answer in Section IV.

- RQ1: What predictive performance can be achieved for different high-level real estate attributes?
- RQ2: How efficient is the fusion of modalities compared to using only single modalities during training?
- RQ3: What is the best fusion strategy to merge the information of the two input modalities?

- RQ4: Are networks trained on complete pairs of photos still capable of correctly predicting missing modality samples?
- RQ5: Does the addition of incomplete data in the training set lead to better test accuracy?

A. Multimodal Network Architectures

The key to multimodal classification lies in the effective fusion of information from different modalities. Therefore, in this work we evaluate the performance of three model architectures that follow different fusion strategies. For all three architectures EfficientNet B0 [24] pretrained on ImageNet [6] is chosen as the backbone architecture to achieve strong classification performance and to allow a fair comparison between all architectures. The three multimodal architecture variants are illustrated in Figure 2 and described in the following.

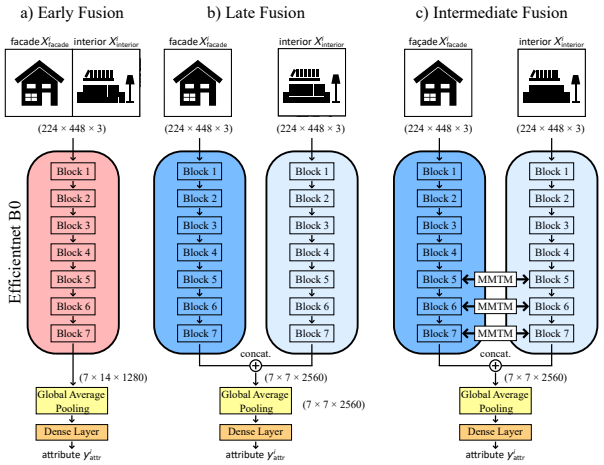


Fig. 2. Overview of the developed network architectures.

Early Fusion: The network architecture in Figure 2 a) represents the concept of early fusion. Both X^i_{facade} and X^i_{interior} of every input pair, each of size (224 x 224 x 3) are horizontally concatenated at the beginning to produce a single input image of size (224 x 448 x 3). The concatenated samples are then fed to the EfficientNet B0 backbone, whose output is a featuremap of size (7 x 14 x 1280). This layer is followed by a global average pooling and a dense layer with softmax activation to output the classification scores.

Late Fusion: Here, instead of concatenating the input images at the beginning, both image modalities are processed in separate subnetworks and are fused at a later stage (Figure 2 b)). Therefore, two separate EfficientNet B0 sub-networks are utilized, which accept input images of size (224 x 224 x 3). In the fusion stage, the two (7 x 7 x 1280) output feature maps are concatenated along the channel dimension and are again processed through a global average pooling layer and a dense layer.

Intermediate Fusion: The third architecture in Figure 2 c) is an extension of the previous one with multimodal transfer modules (MMTM) introduced by Joze et al. [12]. The concept behind multimodal transfer module blocks is illustrated in Figure 3. An MMTM block accepts two feature maps $F_{1,L}$, $F_{2,L}$ from the same Layer L of the two network streams 1 and 2. Within the MMTM block, the information from both feature maps then gets merged through global average pooling and dense layers to generate two gating signals s_1 and s_2. Both gating signals are used to reweight the importance of each featuremap channel of $F_{1,L}$ and $F_{2,L}$. For more details the interested reader can refer to [12]. We use three MMTM blocks, which connect the outputs of the first excitation layers of stages 5, 6 and 7 of EfficientNet B0 [24].

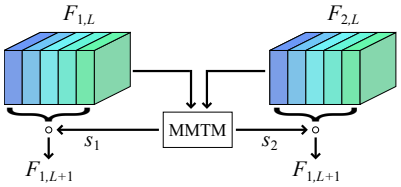


Fig. 3. Concept of multimodal transfer modules (MMTM). $F_{1,L}$, $F_{2,L}$ indicate feature maps of both network streams at layer L. s_1, s_2 are the generated gating signals.

IV. Experimental and Results

In this section, we first provide an overview of the datasets and use cases that serve for the evaluation of our approach. Furthermore, we provide the training details, including the used hyperparameters and the evaluation metrics.

A. Datasets and Use Cases

We evaluate our approach with three different sets of real estate categories and therefore compile the following datasets with respective class labels, making up three different use cases (UC) for evaluation:

- UC1 - Commercial type: classes: industrial, hospitality sector, retail, office
- UC2 - Residential type: classes: apartment, house
- UC3 - Object type: classes: commercial, residential

Each of the respective datasets consists of pairs of facade and indoor photos taken from real estate objects in Austria with corresponding class labels. Often, there are several interior and exterior photos per real estate object. We handle this case by creating multiple unique samples for each real estate object. For example, if six interior and three exterior photos are available for an “office” class commercial object, we create three interior-exterior pair samples of ground-truth class “office” by selecting three random interior photos and assigning one outdoor photo to each. Regardless of whether there are multiple pairs of photos per real estate object, all generated samples are assigned the ground truth class of the associated real estate property.

An overview of these datasets, classes and their partitioning into training, validation and test set can be found in Table I. In our experiments we also want to investigate whether

TABLE I
DATASETS FOR THE THREE INVESTIGATED USE CASES

dataset split	UC1: Commercial type				UC2: Residential type		UC3: Object type	
	industry	hospitality sector	retail	offices	apartment	house	commercial	residential
Train	25 (+30)	30 (+20)	75 (+100)	100 (+50)	300 (+250)	300 (+250)	230 (+200)	600 (+500)
Val	12 (+14)	15 (+10)	37 (+40)	47 (+20)	50 (+50)	50 (+50)	111 (+84)	100 (+100)
Test	14	17	43	50	667	177	124	844

training with additional incomplete data, meaning either indoor X^i_{interior} or facade image X^i_{facade} is missing, can lead to an improvement in prediction accuracy. Therefore, we optionally add incomplete samples to the datasets, where the respective missing visual modality is replaced by a black image. The amount of additional incomplete samples is indicated by the values in parentheses in Table I. When only complete samples are used during training, we refer to the dataset as "*complete*" and when additional missing samples are added we denote it as "*complete + missing*". To avoid bias in favor of one modality, the number of samples with missing facades and missing interior in the "Missing" dataset is kept equal.

B. Training Procedure and Parameters

All experiments are conducted with the following hyper-parameters. Training is performed for a total of 200 epochs with a batch size of 16 and a learning rate of 0.0001 using the Adam optimizer. As a loss function, categorical cross entropy is used. After each epoch, the updated network weights are only saved if the validation loss decreases. To prevent overfitting, we also apply several data augmentation operations including image flipping, rotation, zoom, shear and brightness correction. If an incomplete sample is fed to the network we replace the missing modality with a black image.

C. Evaluation Metric

Since we have a varying amount of data available for each class, our test sets also have different numbers of samples. In our evaluation we nevertheless want to give equal importance to each class and therefore use the Macro F1-score metric, which is defined as follows:

$$\text{Macro F1-score} = \frac{1}{N}\sum_{i=1}^{N}\text{F1-score}_i, \tag{1}$$

where N is the number of classes and i represents the class label.

D. Experiments

In the following, we provide an overview of our experiments. We run experiments for variations of different use cases, modality configurations and multimodal architectures (independent variables). Details on each variable are provided below.

Use Cases: Each experiment is conducted on all three use cases, where each has its corresponding dataset (see Table I).

Modality Configuration: We further want to evaluate whether a multimodal learning approach leads to better results than using only single modality data for training, which is why we also analyze four different modality configurations. The first is the default *complete* configuration, where all data consists of full pairs of interior and facade photos. From this we generate two additional single modality configurations. Specifically, for *facade only* we modify the *complete* configuration by setting all interior photos to black and do the opposite for *interior only*. Finally, we generate a fourth *complete + missing* configuration, in which extra missing modality samples are added to the *complete* configuration (compare Table I).

Multimodal Architecture: We conduct each experiment with all three multimodal network architectures (early fusion, late fusion and intermediate fusion, see Figure 2).

In total this amounts to 36 different experiment configurations (3 use cases, 4 dataset configurations, 3 network architectures). In addition, we repeat every training process three times for each experiment to capture the variations of results originating from different random initializations of the network weights.

V. Results

In the following, we present our experimental results and answer the posed research questions from Section III. The results of all our 36 experiments can be found in Table II. The presented values are Macro F1-scores for the respective test sets, which are additionally averaged over all three training runs. The value inside the parentheses is the standard deviation over all three training repetitions. To evaluate the performance when a network receives samples with missing modality, the same test set is used in three alterations. *test_c* refers to the test set with complete pairs (no missing data). *test_f* and *test_i* refer to the same test set, but here only one modality, facade or interior, is used at a time, while the other one is blackened to simulate missing data in the test sets. For an overview of the split for each modality configuration refer to section IV-A.

With respect to research question 1 (RQ1), Table II shows that the prediction scores differ greatly between the different use cases. While the best Macro F1-score for UC1 (Commercial type) is 0.62, the highest prediction value for UC2 (Residential type) amounts to 0.78. In the UC3 (Object type) setting, the Macro F1-score reaches 0.81. However, it

TABLE II

MACRO F1-SCORES AVERAGED OVER THREE TRAINING RUNS AND IN PARENTHESES THE RESPECTIVE STANDARD DEVIATIONS. RB INDICATES THE RANDOM BASELINE.

Modality Configuration	Multimodal Architecture	UC1: Commercial type (RB = 25%)			UC2: Residential type (RB = 50%)			UC3: Object type (RB = 50%)		
		test_c	test_f	test_i	test_c	test_f	test_i	test_c	test_f	test_i
complete	early	0.54 (0.04)	0.37 (0.05)	0.42 (0.03)	**0.78 (0.01)**	0.76 (0.01)	0.58 (0.01)	0.76 (0.04)	0.71 (0.03)	0.70 (0.03)
	late	0.54 (0.06)	0.36 (0.05)	0.42 (0.03)	0.76 (0.01)	0.76 (0.01)	0.60 (0.01)	0.77 (0.02)	**0.72 (0.01)**	0.71 (0.01)
	intermediate	0.56 (0.05)	0.41 (0.04)	0.44 (0.02)	0.77 (0.01)	0.76 (0.01)	0.61 (0.01)	0.77 (0.02)	**0.72 (0.01)**	0.72 (0.01)
facade only	early		**0.52 (0.03)**			0.72 (0.01)			0.70 (0.02)	
	late		0.40 (0.06)			0.75 (0.01)			0.66 (0.03)	
	intermediate		0.41 (0.03)			**0.77 (0.01)**			0.70 (0.02)	
interior only	early			0.41 (0.04)			0.60 (0.01)			0.69 (0.01)
	late			0.44 (0.02)			0.61 (0.01)			0.72 (0.01)
	intermediate			0.42 (0.05)			**0.62 (0.01)**			0.67 (0.06)
complete+missing	early	0.57 (0.03)	0.40 (0.03)	0.45 (0.03)	0.75 (0.01)	0.72 (0.02)	0.57 (0.01)	0.77 (0.01)	0.71 (0.01)	0.71 (0.03)
	late	**0.62 (0.03)**	0.42 (0.08)	**0.49 (0.02)**	0.75 (0.01)	0.73 (0.01)	0.58 (0.05)	0.79 (0.02)	**0.72 (0.02)**	0.70 (0.02)
	intermediate	**0.62 (0.03)**	0.42 (0.06)	0.48 (0.05)	0.76 (0.01)	0.74 (0.01)	0.55 (0.03)	**0.81 (0.01)**	**0.72 (0.01)**	**0.75 (0.00)**

should be noted that the random baseline (RB) of 50% for UC2 and UC3 is already much bigger than the respective 25% of UC1. Nevertheless, a large margin over the random baseline is achieved for all three use cases.

With research question 2 (RQ2) we wanted to discern whether multimodal learning on both visual modalities is superior to training on individual modalities. For all use cases, *complete* yields better results than *facade only* and *interior only*. There is an increase of 4% of the score for UC1 compared to the best result for the single modality configurations. For UC3, the improvement is 5%. Only for UC2 the performances are almost equal. The reason for the high score for residential properties is probably due to the strong difference in the appearance of facades of apartment buildings and houses, which is also reflected in the similarly high score of the *facade only* configuration. Overall, we can see that training on both modalities provides clear advantages over using only one modality.

Regarding research question 3 (RQ3: which architecture is best suited for multimodal fusion?) we do not reach a clear conclusion. In almost all cases Macro F1-score differences are within 1% or 2%, which does not allow for declaring a clear winner when considering the standard deviations across the three runs. One possible explanation for why the early fusion architecture produces similar results compared to the others, is the fact that both visual modalities concatenated at the input level are RGB images. Hence, the network does not have to deal with information of different dimensionality and domains in its initial layers. It can therefore focus on learning to extract the same low-level features (e.g. edges), which are representative for both input modalities. To summarize the answer to RQ3, we find no significant performance differences between using early, late and intermediate fusion strategies in the evaluated use cases.

Concerning research question 4 (RQ4: generalizability and robustness to missing data) we compare the Macro F1-scores of the *complete* configuration for *test_f* and *test_i* with that of the training configurations *interior only* and *facade only*. Despite the fact that the corresponding networks of *complete* have never been exposed to missing modalities and have only been trained on complete samples they still provide comparable prediction scores for *test_f* and *test_i*. Overall the results show that our multimodal network architectures are capable of handling incomplete input data.

Investigating research question 5 (RQ5) shows that adding additional data with missing modalities leads to better results for two of three use cases. In case of UC1, the increase in Macro F1-score from training on *complete* to *complete + missing* is the largest with almost 6%. For UC2, scores are at the same level, whereas for UC3 performance increases by 6%. These results show that the proposed multimodal network architectures can take benefit of the information contained in the additional incomplete training samples.

VI. QUALITATIVE RESULTS

To further investigate especially the limitations of our approach, we qualitatively analyzed the results. During our experiments, we found that pairs of images that were incorrectly predicted by our networks can be systematically grouped into three main failure types. In this section we want to showcase these failure types using exemplary pairs of photos from our test set and their corresponding predicted labels. For this purpose, we take UC1 (commercial types) and the predictions from the late multimodal architecture for the *complete + missing* modality configuration because it represents one of the most robust combinations. The selected pairs of indoor and facade photos are shown in Figure 4. All pairs are placed in a confusion matrix-like layout, with true positive samples indicated by a green background (diagonal samples). The three failure types are represented by different border colors for the off-diagonal entries.

Unused Clues (blue): This failure type includes samples whose class can be easily recognized by the human observer, but which was not predicted correctly by the network. For

Fig. 4. Confusion matrix with exemplary predicted images from the testset. A Green Background indicates true positive samples. Colored borders indicate different failure types (blue: unused clues, orange: conflicting clues, red: missing clues). Photos taken from justimmo[4] .

example, pair 2) shows two beds in the interior, which is a clear indication for a hospitality object. In addition, in pair 3), the depicted retail property was also misclassified as an industrial building despite having a visible storefront. One explanation for the failed detection in this case could be that in our dataset many industrial buildings have a gray colored floor similar to the one in this pair. In image 9), a lamp post with a brewery logo can be seen, which is a subtle hint for a restaurant that a human observer can understand but was not detected by the network. We hypothesize that this failure type can be mitigated by increasing the total amount of training data available. This way, the network receives more samples from which it can learn relevant patterns.

Conflicting Clues (orange): Some of the samples shown have visual modalities that contain conflicting information. The pair 4) shows photos of an office building with a corresponding looking facade. However, the interior photo depicts a large hall that could also be found in a typical industrial building. The opposite case for an actual industry building can be found in pair 12). Here, the interior photo displays a conference room suggestive of an office building, whereas the exterior resembles an industry building. Pair 14) is a clothing store, which can be recognized by the interior photo. The facade, on the other hand, has nothing in common with typical storefronts. To reduce this failure type, increasing the size of the dataset alone may not be sufficient. In practice, there are often more than two photos available for a given property, all of which could be used in a single model to counteract conflicting modalities. Furthermore, to mitigate such cases, it will be important to assess the representativeness of an image for the target class, i.e., to give less characteristic and speaking images less weight.

Missing Clues (red): The last failure type contains samples that lack any useful clues for classification. In pair 7) a real estate object with an unusual appearance for an office building is shown, which represents a difficult task for our network. Example 11) contains a pair of photos with little useful information. The outdoor photo is a close-up of the door, that gives no hints about the rest of the facade, and the interior photo is a shot of an empty room in suboptimal lighting conditions. A similar issue is present in pair 8). The facade is ambiguous and the room is also empty and lacks information. With respect to this type of failure, the use of additional input photos per property could also be beneficial. In practice, however, we expect that for a certain percentage of real estate objects accurate predictions will fail due to ambiguous or inexpressive pictures. In such cases the incorporation of additional data modalities, e.g. textual descriptions and categorical data can help.

VII. Conclusion

In this paper, we demonstrated the effectiveness and feasibility of using visual data for the prediction of high-level real estate attributes. We leveraged two complementary visual modalities, compared different multimodal fusion strategies and evaluated our approach in three different use cases. Our experiments show that networks trained on both visual modalities (facade and interior) yield better results than networks utilizing only one modality. Furthermore, we could show that our multimodal network architectures provide robust predictions for input samples, which lack one of the two input modalities and that additional training data – even when it is incomplete – can improve the robustness of the models. In future, we plan to extend the proposed multimodal architectures to accept an arbitrary number of input images showing different perspectives of a real estate object.

[4]www.justimmo.at

REFERENCES

[1] E. Ahmed and M. Moustafa, "House price estimation from visual and textual features," *arXiv:1609.08399 [cs]*, Sept. 2016, arXiv: 1609.08399. [Online]. Available: http://arxiv.org/abs/1609.08399

[2] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded Up Robust Features," in *Computer Vision – ECCV 2006*, ser. Lecture Notes in Computer Science, A. Leonardis, H. Bischof, and A. Pinz, Eds. Berlin, Heidelberg: Springer, 2006, pp. 404–417.

[3] J. Bin, B. Gardiner, Z. Liu, and E. Li, "Attention-based multi-modal fusion for improved real estate appraisal: a case study in Los Angeles," *Multimedia Tools and Applications*, vol. 78, no. 22, pp. 31 163–31 184, Nov. 2019. [Online]. Available: https://doi.org/10.1007/s11042-019-07895-5

[4] M. Cain and C. Janssen, "Real estate price prediction under asymmetric loss," *Annals of the Institute of Statistical Mathematics*, vol. 47, no. 3, pp. 401–414, Sept. 1995. [Online]. Available: https://doi.org/10.1007/BF00773391

[5] J.-H. Choi and J.-S. Lee, "EmbraceNet: A robust deep learning architecture for multimodal classification," *Information Fusion*, vol. 51, pp. 259–270, 2019, publisher: Elsevier.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 248–255, iSSN: 1063-6919.

[7] M. Despotovic, D. Koch, S. Leiber, M. Döller, M. Sakeena, and M. Zeppelzauer, "Prediction and analysis of heating energy demand for detached houses by computer vision," *Energy and Buildings*, vol. 193, pp. 29–35, June 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378778818336430

[8] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. Efros, "What makes paris look like paris?" *ACM Transactions on Graphics*, vol. 31, no. 4, 2012.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv:2010.11929 [cs]*, June 2021, arXiv: 2010.11929. [Online]. Available: http://arxiv.org/abs/2010.11929

[10] M. Glodek, S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kächele, M. Schmidt, H. Neumann, G. Palm, and F. Schwenker, "Multiple Classifier Systems for the Classification of Audio-Visual Emotional States," in *Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science, S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Berlin, Heidelberg: Springer, 2011, pp. 359–368.

[11] A. González, G. Villalonga, J. Xu, D. Vázquez, J. Amores, and A. M. López, "Multiview random forest of local experts combining RGB and LIDAR data for pedestrian detection," in *2015 IEEE Intelligent Vehicles Symposium (IV)*, June 2015, pp. 356–361, iSSN: 1931-0587.

[12] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, "Mmtm: Multimodal transfer module for cnn fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 289–13 299.

[13] D. Koch, M. Despotovic, S. Leiber, M. Sakeena, M. Döller, and M. Zeppelzauer, "Real Estate Image Analysis: A Literature Review," *Journal of Real Estate Literature*, vol. 27, no. 2, pp. 269–300, Dec. 2019, publisher: Routledge _eprint: https://doi.org/10.22300/0927-7544.27.2.269. [Online]. Available: https://doi.org/10.22300/0927-7544.27.2.269

[14] Z. Kostic and A. Jevremovic, "What Image Features Boost Housing Market Predictions?" *IEEE Transactions on Multimedia*, vol. 22, no. 7, pp. 1904–1916, July 2020, conference Name: IEEE Transactions on Multimedia.

[15] M. Kächele, P. Thiam, M. Amirian, P. Werner, S. Walter, F. Schwenker, and G. Palm, "Multimodal Data Fusion for Person-Independent, Continuous Estimation of Pain Intensity," in *Engineering Applications of Neural Networks*, ser. Communications in Computer and Information Science, L. Iliadis and C. Jayne, Eds. Cham: Springer International Publishing, 2015, pp. 275–285.

[16] M. Liu and J. Yuan, "Recognizing Human Actions as the Evolution of Pose Estimation Maps," 2018, pp. 1159–1168.

[17] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004. [Online]. Available: https://doi.org/10.1023/B:VISI.0000029664.99615.94

[18] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention Bottlenecks for Multimodal Fusion," *arXiv:2107.00135 [cs]*, June 2021, arXiv: 2107.00135. [Online]. Available: http://arxiv.org/abs/2107.00135

[19] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrušaitis, and L.-P. Morency, "Deep multimodal fusion for persuasiveness prediction," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 284–288.

[20] B. Park and J. K. Bae, "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data," *Expert Systems with Applications*, vol. 42, no. 6, pp. 2928–2934, Apr. 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417414007325

[21] O. Poursaeed, T. Matera, and S. Belongie, "Vision-based real estate price estimation," *Machine Vision and Applications*, vol. 29, no. 4, pp. 667–676, May 2018. [Online]. Available: https://doi.org/10.1007/s00138-018-0922-2

[22] D. Ramachandram and G. W. Taylor, "Deep Multimodal Learning: A Survey on Recent Advances and Trends," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 96–108, Nov. 2017, conference Name: IEEE Signal Processing Magazine.

[23] W. Sun, F. Ma, Y. Li, S.-L. Huang, S. Ni, and L. Zhang, "Semi-Supervised Multimodal Image Translation for Missing Modality Imputation," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 2021, pp. 4320–4324, iSSN: 2379-190X.

[24] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proceedings of the 36th International Conference on Machine Learning*. PMLR, May 2019, pp. 6105–6114, iSSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v97/tan19a.html

[25] L. Tran, X. Liu, J. Zhou, and R. Jin, "Missing Modalities Imputation via Cascaded Residual Autoencoder," 2017, pp. 1405–1414.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[27] Y. Wang, W. Huang, F. Sun, T. Xu, Y. Rong, and J. Huang, "Deep multimodal fusion by channel exchanging," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[28] I.-C. Yeh and T.-K. Hsu, "Building real estate valuation models with comparative approach through case-based reasoning," *Applied Soft Computing*, vol. 65, pp. 260–271, Apr. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494618300358

[29] Q. You, R. Pang, L. Cao, and J. Luo, "Image-Based Appraisal of Real Estate Properties," *IEEE Transactions on Multimedia*, vol. 19, no. 12, pp. 2751–2759, Dec. 2017, conference Name: IEEE Transactions on Multimedia.

[30] M. Zeppelzauer, M. Despotovic, M. Sakeena, D. Koch, and M. Döller, "Automatic Prediction of Building Age from Photographs," in *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, ser. ICMR '18. New York, NY, USA: Association for Computing Machinery, June 2018, pp. 126–134. [Online]. Available: https://doi.org/10.1145/3206025.3206060

A study on robust feature representations for grain density estimates in austenitic steel

Filip Ilic[1], Marc Masana[1,2], Lea Bogensperger[1], Harald Ganster[3] and Thomas Pock[1]

***Abstract*— Modern material sciences and manufacturing techniques allow us to create alloys that help shape our way of living; from jet turbines that withstand extreme stresses to railroad tracks that retain their intended shape. It is therefore an important aspect of quality control to estimate the microstructural properties of steel during and after the manufacturing process, as these microstructures determine the mechanical properties of steel. This estimation has for a long time been a labor intensive and non-trivial task which requires years of expertise.**

We show that modern deep neural networks can be used to estimate the grain density of austenitic steel, while also applying a visualization technique adapted to our task to allow for the visual inspection of why certain decisions were made. We compare classification and regression models for this specific task, and show that the learned feature representations are vastly different, which might have implications for other tasks that can be solved via discretization into a classification problem or treating it as an estimation of a continuous variable.

I. INTRODUCTION

Not all steel is created equally. Other than the ratio of carbon and other metals that are used in the alloy when it is being forged to steel, different modes of cooling, heating, and hardening produce variations in steel. Broadly speaking, steel can be classified into austenite, martensite, and under certain circumstances even a mixture of both. Martensite forms when steel is quenched very quickly, whereas austenite forms through a lengthy cooling process. Even within the austenite cooling process, there are many factors that influence the development of microstructures within the steel that contribute to the graining process, i.e. the formation of individual grains. Determining the characteristic grain size of the sample, which is used to determine the grain density, is important for many applications as it relates to the tensile and compressive stresses that the material is able to withstand. These grains and other microstructures of the resulting steel can - through an extensive etching and cleaning process - be made visible under a light microscope [10].

Traditionally, austenitic steel grain density is estimated by costly and labour intensive work done by a metalographer where etched steel samples are manually inspected under a light microscope. Currently the most reliable way to perform this grain density estimate is by including a template, that is projected onto the viewfinder of the microscope. The metalographer then uses this template to determine the grain density by comparing it to the different available templates. Since only a 2-dimensional cross-section of the 3-dimensional material is visible, grains might appear smaller or larger than the average grain size within the material due to the slicing process. It is therefore a requirement that a judgement is made based on the relative distributions of grain density within a single slice of the sample.

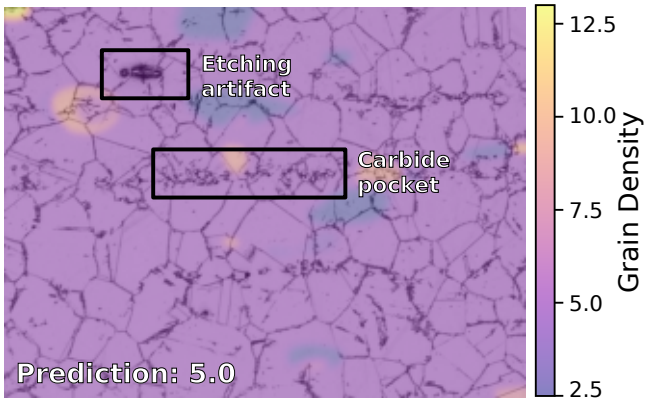


Fig. 1. A single image of an austenitic steel sample taken with a light microscope at 100-fold magnification. Our method estimated the overall density of the sample to be 5.0. The overlayed heatmap is created with a proposed visualization scheme, detailed in Section VI, that aids in understanding the decision process of the network, as a pixel-wise density estimate can show inhomogeneous regions if they are present. Note that the learned representation is robust enough to ignore sample preparation artifacts and carbide pockets that have a similar appearance to higher density grain regions.

In this paper, we propose a deep learning-based approach to estimate austenitic steel grain density from a single image. We explore how classical cross-entropy-based losses allow to learn classification models with state-of-the-art performance. However, we find that classification models - at least in the domain of grain density estimation - come at a price when comparing it to similar, albeit slightly less performant regression models, that show more resilience when dealing with out of distribution samples, and appear to have a more robust and human interpretable feature space. We therefore also propose to use regression-based losses that are capable of predicting a continuous grain density, at the cost of a slight decrease in performance.

It is notoriously difficult to explain the decision making process of deep neural networks, which can often be a source of confusion when applied and deployed in real world applications. It is therefore important, to provide tools to visualize the model's decisions, and understand the failure cases and the reasons for a failure. Recently, some methods

[1] Graz University of Technology, Austria
[2] Silicon Austria Labs, TU-Graz SAL DES Lab, Austria
[3] JOANNEUM RESEARCH Forschungsgesellschaft mbH, Austria
Corresponding Author: filip.ilic@icg.tugraz.at

have been proposed to evaluate the confidence by using class activation maps [32]. We adapt a well known algorithm – GradCAM [27] – that allows us to visualize local regions within a single sample, that can a) show us glimpses of the underlying feature embeddings and whether they really encode relevant information that trained metalographers would look for, and b) give more fine grained analysis of the input image than just the classification label, shown in Fig. 1.

A challenge of using deep learning models in this domain is that deep neural networks require large amounts of data to reach a satisfying degree of robustness. Because the data acquisition and labeling of steel samples requires thorough metalographic knowledge, this data is rather scarce. Therefore, we propose a heavy data augmentation scheme that allows to generate grain densities of continuous granularity, even when only whole grain (i.e. 4.0, 5.0, etc.) austenite data is available, as it often is.

In summary, our contributions are the application of classification and regression based deep learning models to the domain of microstructural analysis of austenite steel, with a focus on the differences in interpretability of the resulting feature representation that the two modes of learning yield. Furthermore, we propose a data augmentation scheme which could be extended to other datasets that are fractal-like or display self-similarity. We show that its usage improves performance across a variety of different models. We present through ablations on classification and regression models that in general classifiers perform better than regressors in this setting. However, this improved performance comes at the cost of a decrease in robustness and interpretability.

II. RELATED WORK

In the past, many insights into material composition and corresponding material properties were derived from expert knowledge and experience. Nowadays, data generated by simulations and measurement systems are becoming more available, thus moving away from physically-based tests. Agrawal and Choudhary [1] introduce the term *deep materials informatics* in the context of data-driven technologies and provide a comprehensive overview of challenges and applications of deep learning with respect to learning chemical compositions of materials, prediction of crystalline structures, (3D) microstructure analysis, and microstructure reconstruction [6]. Furthermore, [12] illustrate opportunities and current paths, where machine learning will have significant influence on material science.

Automated detection or classification of microstructures is the central theme of metallographic studies. Chowdhury et al [7] use image analysis and machine learning to discriminate whether samples have dendritic morphologies or not. DeCost and Holm [13] use a feature-based approach to identify generic signatures of microstructures. These serve as the basis for a Support Vector Machine (SVM) [8] classifier to distinguish 7 microstructure classes. Similarly, Gola et al [17] employ an SVM model for reproducible and objective microstructure classification and achieve classification accuracy greater than 90% for cast iron samples. The morphological data comes from both optical microscopy and electron microscopy images, and the mixed microstructure exhibits a variety of graphite morphologies. An extension of the classification system to deep learning techniques achieved 95% accuracy on unprocessed electron micrographs of low-alloy steels [3], [25]. Here, a combination of CIFARNet, a modification of LeNet [22], and a pretrained VGG16 network [28] were used. Mulewicz et al. [23], [24] distinguish 8 classes of microstructures of different steel grades (C15, C45, C60, C80, V33, X70, and non-hardened steel) from optical microscopy images with the aid of a deep network structure based on ResNet18 [18]. The authors of [15] train models with U-Net architectures with about 30-50 micrograph samples in order to achieve robust segmentation for bainite microstructures. To segment microstructures into four relevant domains ("grain boundary carbide, spheroidized particle matrix, particle-free grain boundary denuded zone, and Widmanstätten cementite"), DeCost et al [11] use pixel-based machine learning [4]. Their segmentation model was compared to the results of microscopic annotation by metalographer using 24 carbon steel samples. Although direct comparison in microstructures (< 5 pixels) was not possible and demonstrated the need for high quality training data, it was still possible to show the effectiveness of deep learning in the analysis of complex microstructures. Albuquerque et al. [9] apply a multilayer perceptron with backpropagation to achieve a microstructure segmentation for cast iron images. Verification on a test set of 60 images showed high correlation to human ground truth. In this line Bulgarevich et al. [5] apply a Random Forest classifier to optical microscopic images of steels for an automated segmentation. Austenite grain density is a significant variable in the AI system of Kuziak [21], which allows the estimation of different phase constituents occurring during the cooling process.

III. DATASET

To perform a density analysis the grains within the steel need to be made visible. Various types of acids are used which etch the weak spots of the metal surface, i.e. the grain boundaries or other impurities of the metal, away first, leaving behind a darkened appearance. The prevailing industry standard to measure the grain density within the material is the ASTM e112 [2] norm. It specifies a 100-fold magnification at which the optical microscope images are captured. Therefore all our images are taken with a 100-fold magnification, and in total consist of 242 images that have a resolution of 1280×960. We split them into 125 train, 53 validation, and 64 test images, keeping the distribution of classes balanced.

The dataset contains images from whole-grade densities ranging from 4.0 to 13.0 with increments of 1.0, and additionally the grain density 2.5. This range of grain densities are provided by the manufacturing process at the steel mill. Fig. 2 shows austenitic steel with various grain densities; it also shows the variation in appearance that is due to the different alloys, and variations in the etching process.

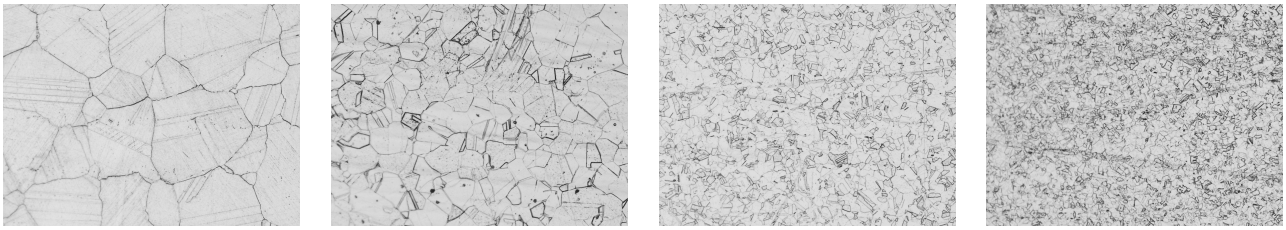


Fig. 2. Samples of varying grain densities. The grain density increases from left to right and exhibits fractal like self similarity at different scales. It does however produce a variety of artifacts due to the etching process depending on size of individual grains and variation among samples. While the grain density is fundamentally a continuum, it is often discretized to whole- or half-grades in practice.

IV. GRAIN DENSITY ESTIMATION

The problem of grain density estimation can be framed as an image classification task, with the increased complexity that naturally occurring grain density variations might not be homogeneously distributed across the entire sample. Image classification has seen a massive shift from hand-crafted feature detectors towards the use of different deep learning techniques. When data is limited, a common technique is to pretrain on a large dataset and only fine-tune the network to the specific domain. This exploits a good initialization of the network parameters to learn an adjusted representation of that smaller domain [26]. Since the amount of annotated data which contains information about microstructures, including grain density, is usually limited due to its acquisition cost, we propose to use fine-tuning with a cross-entropy loss on a pretrained classification network. To the best of our knowledge, this popular technique has not been applied to this setting. The closest work which uses deep learning for microstructural analysis in steel is [3]. Although this work is not applied to austenitic steel, nor evaluated with grain density estimates, we consider it in our comparisons.

Classification allows for the network to learn representations which project the input data into a feature space where different classes can be easily discriminated without any specific ordering. However, due to the nature of steel grains spanning a continuum of different sizes, we also consider to frame the grain density estimation as a regression problem. This allows the network to not only discriminate between different classes, but also maps to a feature space that implicitly preserves grain density order.

Classification. We consider a backbone pretrained feature extractor Φ parameterized by weights θ_Φ and a classifier Ψ parameterized by weights θ_Ψ. We define $\mathbf{o}(\mathbf{x}) = \Psi(\Phi(\mathbf{x};\theta_\Phi);\theta_\Psi)$ as the output logits of the network given an image $\mathbf{x}$. Then, given $\mathbf{y}$ as the one-hot encoding of the ground truth label corresponding to the N classes (grain densities), we consider the cross-entropy loss

$$\mathscr{L}_{\text{CE}}(\mathbf{x},\mathbf{y};\theta_\Phi,\theta_\Psi) = \sum_{k=1}^{N} y_k \log \frac{\exp(\mathbf{o}_k)}{\sum_{i=1}^{N}\exp(\mathbf{o}_i)}. \tag{1}$$

Regression. We use the same feature extractor and head as in classification, together with output logit $\mathbf{o}(\mathbf{x})$ given an image $\mathbf{x}$. However, given $\mathbf{y}$ as the actual numerical value of the ground truth grain density, and $\mathbf{d} = \mathbf{o}(\mathbf{x}) - \mathbf{y}$, we define the regression loss as a smooth ℓ_1 loss

$$\mathscr{L}_{\text{S1}}(\mathbf{x},\mathbf{y};\theta_\Phi,\theta_\Psi) = \begin{cases} \frac{\mathbf{d}^2}{2\alpha}, & \text{if } |\mathbf{d}| < \alpha \\ |\mathbf{d}| - \frac{\alpha}{2}, & \text{otherwise} \end{cases}, \tag{2}$$

where $\alpha = 1$. This threshold α specifies when the loss function changes between ℓ_1 and ℓ_2^2. This loss is less sensitive to outliers, than the mean squared error and can help prevent exploding gradients [16].

Data augmentation. As stated earlier, austenitic steel data for microstructural analysis is costly to acquire and difficult to annotate correctly. This leads to generally small datasets, which can be an issue for deep learning models. However, apart from fine-tuning on pretrained models, another popular training strategy is data augmentation, which consists of altering and extending samples from the dataset with class preserving transformations. The transformed samples increase the number of images to be learned from and help the model generalize better, and to have a more robust representation of the target domain.

The grain density G is determined by $N = 2^{G-1}$, where N is the number of grains per square inch at $100\times$ magnification. The different grain densities exhibit similar structures and patterns at different scales with self-similar features. Therefore, various magnifications of samples with their corresponding adapted labels can be generated from image patches to simulate larger or smaller grain densities by cropping and resizing them in accordance with the grain density formula. Our proposed data augmentation strategy consists of generating new samples which differ at a maximum of ± 0.5 grades from the original. In the case of classification this is set to a binary ± 1.0 to align with our class labels. In addition we perform the common data augmentation best practices: random rotations between 0°and 360°, horizontal and vertical flips, and contrast jitter to simulate possible changes in the lighting conditions during data acquisition or variations in the etching strength during sample preparation. In the experimental sections we will denote the additional rescaling during data augmentation as $\lambda(\cdot)$, and apply the rest of mentioned transformations to all reported experiments.

Image and crop augmentation. Regression or classification can be performed by passing the whole image or crops of a fixed size to the model. Our proposed data augmentation

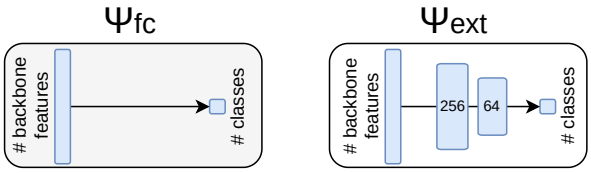


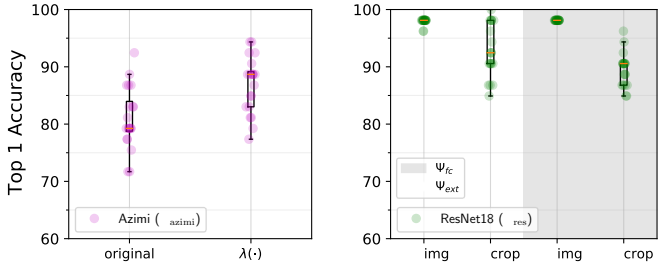
Fig. 3. Two different proposed heads for image classification: Ψ_{fc} is defined as a single fully-connected layer, whereas Ψ_{ext} is defined with an extension of 2 more intermediate fully-connected layers. When doing regression, the last layer is replaced by a single output.

is nearly identical for both of these scenarios with one small difference. In the case of whole image augmentation, the re-scaling function $\lambda(\cdot)$ does not operate within the bounds of ± 0.5, but between 0.0 and -0.5 – analogously 1.0 for classification. This means that we only generate samples with a smaller grain density. This is because the re-scaling to a higher grain density would require generating or replicating new image regions to fit the space left empty from the re-sizing. Using the whole image will be denoted as *img*, while the use of crops of size 224×224 will be denoted as *crop*.

Architectures. Due to the relatively small size of the dataset, retraining state-of-the-art feature extractors such as ResNet18(Φ_{res}) [18] or AlexNet(Φ_{alex}) [20] architectures from scratch yields worse results than using pretrained models. Therefore, we use pretrained Φ_{res} and Φ_{alex} models on Imagenet [14] as the backbones in our experiments. These two architectures have shown to perform well in different image classification and regression tasks, some of which share the domain of microstructure analysis [3]. The two architectures also represent two paradigms in deep learning; convolutions alone, or incorporating residual blocks. Furthermore, we propose to use two different heads applied on top of the feature extractor: Ψ_{fc} and Ψ_{ext} (see Fig. 3). Ψ_{fc} is a single fully-connected layer on top of the feature extractor, commonly used in fine-tuning from a pretrained model. The other, Ψ_{ext} is an extended head with two intermediate fully-connected layers, to allow for a larger capacity in the classification or regression head.

Metrics. We evaluate image classification performance with Top 1 accuracy. However, for regression, exact prediction of the grade is neither necessary nor effective. A more comparable metric to classification is to allow for a margin of ± 0.5 around the regressed prediction. If the prediction lies within the margin we still consider it to be correct. This relates to the available metalographic data being labeled either in whole-grain or sometimes in half-grain steps.

Experimental setup. Each network architecture is trained with Adam [19] with an initial learning rate of 3e-4. Training spans 1,500 epochs and the final model is chosen from the epoch with the lowest validation loss before evaluating on the test split. Each experiment consists of 20 seeds to measure the robustness to different initializations.



a) Azimi et al. [3] b) Ours

Fig. 4. Grain density estimation with classification. We compare Azimi et al. [3] (purple) and our proposed ResNet18-based architecture (green) with various configurations. We also demonstrate the effectiveness of our data augmentation $\lambda(\text{crop})$ on [3].

V. EXPERIMENTAL RESULTS

To assess the performance of the proposed strategies, we first compare results on classification, then on regression, and finally we summarize and discuss them together.

Classifying Grain Density. The first approach we consider is using classification networks to solve the problem of grain density estimation. We show our results across the different configurations introduced in Section IV and compare them in Fig. 4b). Furthermore we compare our models directly to the approach proposed in [3] (Φ_{azimi}). We note that our best configuration Φ_{res}^{cls} outperforms Φ_{azimi} by 19.4% on average, if Φ_{azimi} is trained with their proposed scheme. However, if we employ our proposed data augmentation pipeline $\lambda(\text{crop})$ on Φ_{azimi} performance improves and the gap is reduced to 11.3%, yielding an improvement of 8.1% just by using $\lambda(\cdot)$. As Φ_{alex} never exceeds 60% Top 1 accuracy across the various settings it is omitted from the ablation figure.

Regarding our results of Φ_{res}, we show that training on whole images results in better performance than training on image crops. The network heads Ψ show no effect when training on whole images, and a slight increase of performance when using Ψ_{ext} on crops.

Regressing Grain Density. We also investigate using regression networks to estimate the grain density. In Fig. 5, we show an ablation of the regression configurations. Φ_{res} outperforms Φ_{alex} in every configuration that is comparable. This is especially impressive as Φ_{res} has only roughly 11 million parameters, whereas Φ_{alex} has around 60 million parameters. It is easy to conclude that there is neither a gain in performance nor a gain in computational cost in using Φ_{alex}. We find that the best performing model is Φ_{res} with a plain Ψ_{fc} head, using image crops and our λ augmentation. This is also shown and summarized in Table I.

Regarding the heads, Ψ_{fc} in combination with Φ_{res} yields models that have a smaller standard deviation. This is explained by the fact that heads with more capacity tend to over-fit on the limited data, while the pretrained backbone is robust enough to not degenerate. Another interesting finding is that passing the whole image (i.e. global information) through the network generally performs worse across all tested configurations than using crops, except for one outlier. This indicates that local information plays a more important

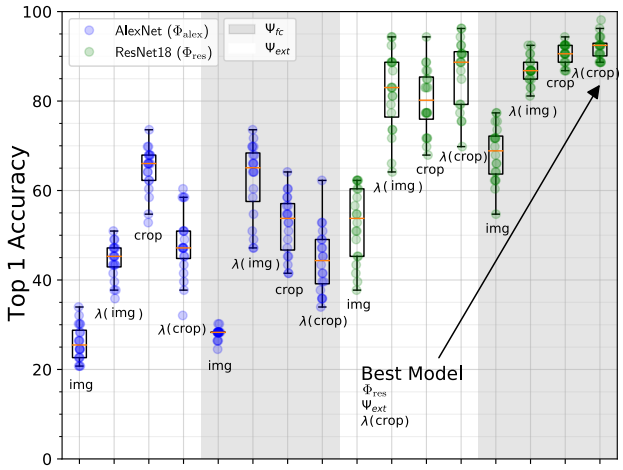


Fig. 5. Grain density estimation with regression. Comparison of Φ_{res} (green) and Φ_{alex} (blue) when training on images (img) or crops (crop), combined with classifiers Ψ_{fc} (gray) and Ψ_{ext} (white), and with data augmentation $\lambda(\cdot)$.

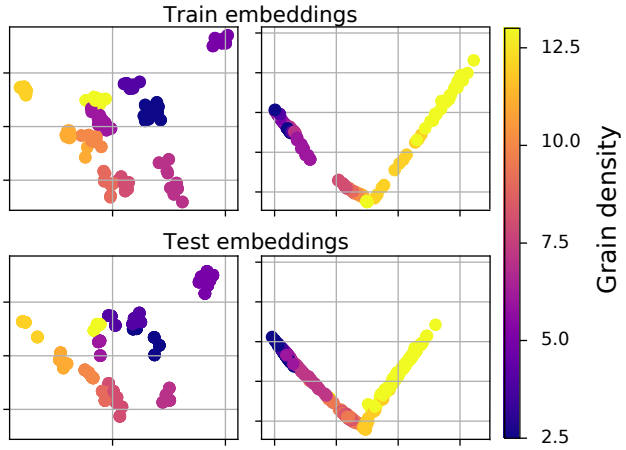


Fig. 6. Classifier Φ_{res}^{cls}(left) and Regressor Φ_{res}^{reg}(right) feature space visualization with PCA. While classification models outperform regression models w.r.t. Top 1 accuracy, it might come at a cost. The learned feature representation of the classifier, while good at separating classes, does not span the grain density space continuously according to their size. This is in contrast to regressors that clearly show a grain-density axis.

TABLE I

SUMMARY OF GRAIN DENSITY ESTIMATION

	Classification		**Regression**	
	Crop	Image	Crop	Image
$\Phi_{\text{res}}, \Psi_{\text{ext}}$	89.91 ± 3.02	98.11 ± 0.09	86.42 ± 7.66	81.98 ± 8.80
$\Phi_{\text{res}}, \Psi_{\text{fc}}$	93.02 ± 4.07	97.92 ± 0.58	91.89 ± 2.68	86.98 ± 2.99

TABLE II

MEAN ABSOLUTE ERROR ON CLASSES OF UNSEEN GRAIN DENSITIES

Model	Train	Test	Unseen Classes
Best Classifier $(\Phi_{\text{res}}, \Psi_{\text{ext}}^{cls})$	0.000	0.019	1.038
Best Regressor $(\Phi_{\text{res}}, \Psi_{\text{fc}}^{reg})$	0.135	0.216	**0.646**

role, and that the anticipated inhomogeneities within a sample do not contribute to wrong estimates, which is surprising because Classifiers Φ_{res}^{cls} all performed considerably better with global than with local information. Finally, our data augmentation strategy $\lambda(\cdot)$ increases performance by $\sim$7% on average w.r.t. Φ_{res}. The best model configuration is a combination of Φ_{res}, Ψ_{fc}, and $\lambda(\text{crop})$, as seen in Table I.

Discussion We generally observe that classification models outperform their regression counterparts (see Table I). In contrast to regression which prefers crops to images, we find that classifiers exhibit preference towards whole images. This already hints that the learned feature representation for regression and classification is drastically different, which we explore further in the following section.

VI. FEATURE REPRESENTATION AND VISUALIZATION

Interpretability. Visualizing the feature space of learned image representations is often done to gain insight into the decision making process. When visualizing embeddings $\Phi(\mathbf{x}) \in \mathbb{R}^{256}$ from image $\mathbf{x}$, we need to reduce its high dimensionality to allow for better visual analysis. This step could be done with methods such as Principal Component Analysis (PCA) [30] or t-Stochastic Neighbor Embedding (t-SNE) [29]. We choose PCA since distances in the projection are preserved, unlike in non-linear projections such as t-SNE.

We forward pass our training samples through Φ_{res}^{cls} and Φ_{res}^{reg}, with *cls* and *reg* denoting the best classifier and regressor network backbones. We then apply the same projection to the test set and observe where they end up in feature space. Results are shown in Fig. 6. A drastic difference between the feature representations of Φ_{res}^{cls} and Φ_{res}^{reg} can be observed. The embedding space of the classifier does not arrange classes corresponding to grain densities in any particular order. Instead, classes form clusters where interpolation in the feature space does not equal interpolation in grain density. Contrarily, a very orderly arrangement of grain densities emerges when learning with regression, as shown in the right column of Fig. 6. These results are particularly interesting as we previously show that Φ_{res}^{cls} outperforms Φ_{res}^{reg} by a significant margin, thus one could relate a more structured feature space representation to better performance.

Out-of-distribution robustness. In order to investigate if the ordered grain density feature structures emerging from learning with a regressor is beneficial, we explore inference on unseen and out-of-distribution data that *does* occur in real world scenarios. We further have metalographers annotate 668 new samples belonging to half-grade density austenite steel – which finer partition is commonly used in real world applications – and captured with a similar setup as the data described in Sec. III. Concretely this new dataset consists of austenite steel images with classes corresponding to grain densities ranging from 3.5 to 12.5, in increments of 1.0 – with only the grain density 10.5 missing.

In Fig. 7, we show the embedding plots of these unseen classes, both for Φ_{res}^{cls} and Φ_{res}^{reg}. Once more it can be observed

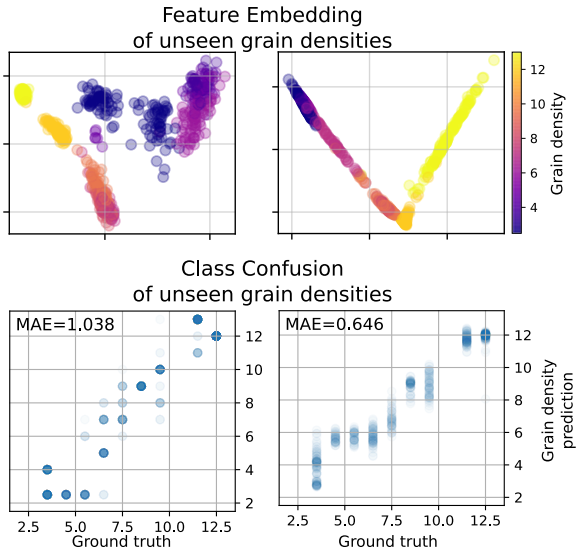


Fig. 7. Feature embeddings of unseen grain densities on a classification (left) and regression (right) model. Bottom row shows the regression plot, in essence a confusion matrix, where each sample is plotted relating its ground truth and predicted value. The Mean Absolute Error (MAE) is considerably lower for the regression model.

that the feature space exhibits mostly a continuous representation of the grain densities in the case of Φ_{res}^{reg}, and Φ_{res}^{cls} exhibits the same clustering behaviour. Not only is this shown qualitatively in the visualization, but is also quantitatively established in terms of Mean Absolute Error (MAE) over the out-of-distribution samples. In conclusion, Table II summarizes our findings by showing that the classifier performs both better on train and test sets, but generalizes worse to out-of-distribution samples. In contrast, regression presents a potential trade-off between the performance of a model and its interpretability at the feature representation, which allows evaluation of intermediate grain densities without re-training.

Grain density attention mapping. Work that focuses on visualization and explainability of convolutional neural networks has been around almost since their inception [31]. A common technique, especially for classification-based methods, is the use of class activation map (CAM) [32] algorithms. Since the grain density estimation is also framed as a classification problem, we can apply GradCAM [27], a popular CAM algorithm, to highlight areas in images that correspond to particular classes. We exploit the fact that an ordering of the grain density classes exists, which enables us to analyse image structures that lead to high activations in the output neurons. This can be used in order to visually perform a grain density homogeneity estimation.

GradCAM generates attention maps based on the gradients of a network w.r.t. a particular class and image. We perform a GradCAM step for every single class given, stack the generated attention maps, and compute the maximally activated class value for each pixel. The resulting scalar field is a pixel-wise class activated discriminative map.

To test the robustness and predictive capabilities of our proposed architectures we splice together an image consisting of 4 individual images of different grain densities. The images used correspond to those in Fig. 2. The resulting class attention maps are shown in Fig. 8, for a model trained on whole images and one trained on crops. We observe that the classifier network that was trained on whole images has difficulties to detect the boundaries of the various grain densities, whereas the network trained on crops shows no such limitation and produces a heatmap delineating the spliced quadrants very well. The crop trained model processes the individual crops separately, which are then assembled to a single attention map. The inhomogeneity detection and visualisation provided by the crop-trained model could be explored in future work, because the homogeneity of austenitic steel is useful for determining its mechanical properties.

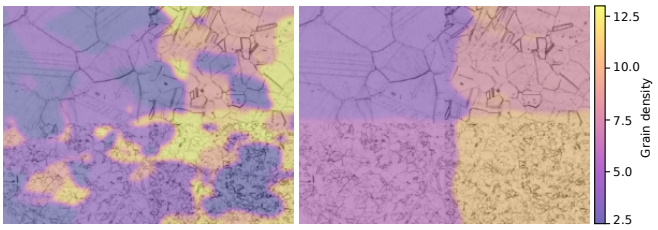


Fig. 8. An artificially spliced image from 4 different grain densities trained with whole images (left) and crops (right). The overlayed heatmap is generated by our proposed argmax GradCAM modification to provide pixel-wise grain density estimates. Best viewed digitally.

VII. CONCLUSION

We explore classification and regression with deep neural networks for estimating the grain density of austenitic steel samples taken with optical microscopes. We show that classification models overall yield better results than comparable regression models. Our findings show that the learned feature representation of classifiers and regressors differs drastically. The feature embedding of regressors yields an interpretable axis that corresponds to the actual grain density, whereas classifiers do not seem to encode the grain density as a major dimension in their feature space, and instead partition it into rigid, easily separable clusters. This is also reflected in the results that compare the performance of both types models on previously unseen grain density samples. Dealing with such out-of-distribution samples is especially important in the context of real-world applications. Since regression is shown to be robust w.r.t. out-of-distribution samples while maintaining accurate grain density estimates, we demonstrate a feasible way of additional quality control in steel mills. We also show the adaptation of a popular CAM algorithm to visualize grain densities and inhomogeneities within a sample, which also provides insight into the learned feature representation. Due to limited data, common in these settings, we introduce a novel data augmentation technique tailored to grain density estimation, which is shown to improve the performance of both classifiers and regressors.

ACKNOWLEDGMENT

This work was supported by *Land Steiermark* within the research initiative "Digital Material Valley Styria". Marc Masana acknowledges the support by the "University SAL Labs" initiative of *Silicon Austria Labs (SAL)*.

REFERENCES

[1] A. Agrawal and A. Choudhary, "Deep materials informatics: Applications of deep learning in materials science," *MRS Communications*, vol. 9, no. 3, pp. 779–792, 2019.

[2] E.-. ASTM, "Standard test methods for determining average grain size," *ASTM International: West Conshohocken, PA, USA*, 2004.

[3] S. Azimi, D. Britz, M. Engstler, M. Fritz, and F. Mücklich, "Advanced Steel Microstructural Classification by Deep Learning Methods," *Scientific Reports*, vol. 8, 02 2018.

[4] A. Bansal, X. Chen, B. Russell, A. Gupta, and D. Ramanan, "Pixelnet: Representation of the pixels, by the pixels, and for the pixels," *arXiv preprint arXiv:1702.06506*, 2017.

[5] D. Bulgarevich, S. Tsukamoto, T. Kasuya, M. Demura, and M. Watanabe, "Pattern recognition with machine learning on optical microscopy images of typical metallurgical microstructures," *Scientific Reports*, vol. 8, 12 2018.

[6] K. Choudhary, B. DeCost, C. Chen, A. Jain, F. Tavazza, R. Cohn, C. WooPark, A. Choudhary, A. Agrawal, S. J. L. Billinge, E. Holm, S. P. Ong, and C. Wolverton, "Recent advances and applications of deep learning methods in materials science," 2021.

[7] A. Chowdhury, E. Kautz, B. Yener, and D. Lewis, "Image driven machine learning methods for microstructure recognition," *Computational Materials Science*, vol. 123, pp. 176 – 187, 2016.

[8] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[9] V. H. C. de Albuquerque, P. C. Cortez, A. R. de Alexandria, and J. M. R. Tavares, "A new solution for automatic microstructures analysis from images based on a backpropagation artificial neural network," *Nondestructive Testing and Evaluation*, vol. 23, no. 4, pp. 273–283, 2008.

[10] C. G. de Andrés, F. Caballero, C. Capdevila, and D. San Martın, "Revealing austenite grain boundaries by thermal etching: advantages and disadvantages," *Materials Characterization*, vol. 49, no. 2, pp. 121–127, 2002.

[11] B. L. DeCost, T. Francis, and E. A. Holm, "High throughput quantitative metallography for complex microstructures using deep learning: A case study in ultrahigh carbon steel," *Microscopy and microanalysis: the official journal of Microscopy Society of America, Microbeam Analysis Society, Microscopical Society of Canada*, vol. 25 1, pp. 21–29, 2018.

[12] B. L. DeCost, J. R. Hattrick-Simpers, Z. Trautt, A. G. Kusne, E. Campo, and M. L. Green, "Scientific ai in materials science: a path to a sustainable and scalable paradigm," *Machine learning: science and technology*, vol. 1, no. 3, p. 033001, 2020.

[13] B. L. DeCost and E. A. Holm, "A computer vision approach for automated analysis and classification of microstructural image data," *Computational Materials Science*, vol. 110, pp. 126 – 133, 2015.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 248–255.

[15] A. I. Durmaz, M. Müller, B. Lei, A. Thomas, D. Britz, E. Holm, C. Eberl, F. Mücklich, and P. Gumbsch, "A Deep Learning Approach for Complex Microstructure Inference," *Nature communications*, vol. 12, no. 1, 2021.

[16] R. Girshick, "Fast r-cnn," in *International Conference on Computer Vision*. IEEE, 2015, pp. 1440–1448.

[17] J. Gola, D. Britz, and F. Mücklich, "3D-Gefügeforschung und neue Möglichkeiten der zuverlässigen Gefügeklassifizierung durch Kombination mit maschinellem Lernen," *METALL*, vol. 72, pp. 454–456, 2018.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 770–778.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[21] J. Kusiak and R. Kuziak, "Modelling of microstructure and mechanical properties of steel using the artificial neural network," *Journal of Materials Processing Technology*, vol. 127, pp. 115–121, 09 2002.

[22] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, vol. 86, no. 11, 1998, pp. 2278–2324.

[23] B. Mulewicz, G. Korpala, J. Kusiak, and U.Prahl, "Deep convolution neural networks in classification of metals microstructure," in *International Conference on Adaptive Modeling and Simulation, ADMOS 2019*, 2019.

[24] B. Mulewicz, G. Korpala, J. Kusiak, and U. Prahl, "Autonomous interpretation of the microstructure of steels and special alloys," *Materials Science Forum*, vol. 949, pp. 24–31, 03 2019.

[25] M. Müller, D. Britz, and F. Mücklich, "Application of trainable segmentation to microstructural images using low-alloy steels as an example," *Practical Metallography*, vol. 57, pp. 337–358, 04 2020.

[26] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 1717–1724.

[27] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *International Conference on Computer Vision*. IEEE, 2017, pp. 618–626.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*. IEEE, 2015.

[29] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[30] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[31] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*. Springer, 2014, pp. 818–833.

[32] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 2921–2929.

On the Influence of Beta Cell Granule Counting for Classification in Type 1 Diabetes

Lea Bogensperger[1], Marc Masana[1,2], Filip Ilic[1], Dagmar Kolb[3,4], Thomas R. Pieber[4,5], and Thomas Pock[1]

***Abstract*— Patients suffering with type 1 diabetes show a major reduction of β-cells within their pancreas. By analyzing tissue samples containing granules – the insulin-producing units within the β-cells – we aim to gather more information on the respective healthy and diabetic phenotypes, which could lead to further understanding the pathogenesis of the disease. To this end, we use a deep learning approach to investigate whether assumptions on the pathological status can be made based on electron micrograph images of β-cells. To support the decision-making process we explore whether estimating the number of granules can be used to aid in discriminating healthy from diabetic samples. Furthermore, we demonstrate that multi-task and transfer learning strategies can lead to more accurate predictions. Finally, this work intends to contribute to a more in-depth understanding of the structural mechanisms in type 1 diabetes, which is essential to design better approaches to a tailored treatment.**

I. INTRODUCTION

Type 1 diabetes is an autoimmune disease that is characterized by the destruction of insulin-producing β-cells [2]. The body's immune system attacks its own essential insulin-producing mechanism within the pancreas thus impeding a normal blood glucose regulation after food intake. The main participants in this process are the granules of β-cells that can be found within the islets of Langerhans in the endocrine pancreas. They are composed of an insulin-producing core and surrounded by a less dense halo, as explained in [11]. The author further describes how the granules are suspected to be the focus of the autoimmune-mediated β-cell destruction due to immunogenic targets. An example of such granules is shown in Figure 1, where part of a healthy and a non-obese diabetic (NOD) mouse are shown. Following the attack of the immune system, the overall number of granules contained in the β-cells gets drastically decreased and the remaining β-cells are stressed in trying to maintain insulin supply [3].

By visually inspecting electron micrograph slices such as in Figure 1, it is at first glance not trivial to discriminate the microscopic images between healthy and NOD. Therefore, an important research question deals with finding features that allow for distinguishing between β-cells in healthy and NOD mice. The number of granules that can be found within β-cells seems to be a strong indicator. Furthermore, there are hypotheses in the medical domain regarding structural changes of the entire β-cell [13] which can result in changed appearance of the granules regarding shape and size.

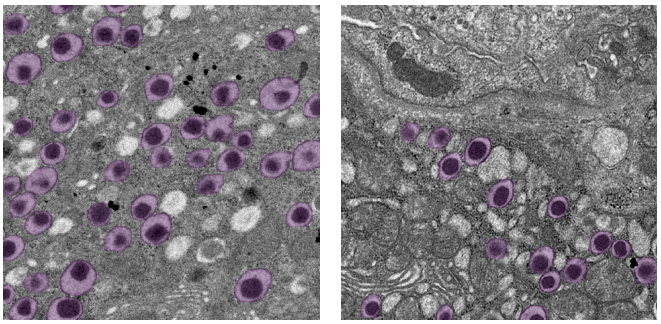


Fig. 1. Slices revealing β-cells of a healthy (left) and a diabetic NOD (right) mouse. Numerous granules, colored in purple, are visible in both samples. They are characterized by their circular structure containing a dark insulin-producing core surrounded by its vesicle characterized by its typical white halo.

We therefore propose a deep learning based approach to discriminate between slices containing β-cell granules of healthy and NOD mice. Our focus is to learn a classification system that can distinguish between healthy and early stage type 1 diabetes in electron micrograph images of β-cells with their insulin-producing granules. Further, fueled by the interesting findings that the number of granules seems to be able to distinguish to an extent between healthy and NOD slices (see Section IV), we pursue to train a granule-counting system that for a given β-cell slice learns to count the number of granules present in the sample.

To summarize, our main contributions are:

- we demonstrate the applicability of a deep learning based approach in classifying electron micrograph images as either healthy or NOD,
- we show that a network can also be trained to estimate the number of insulin-producing granules in a given input image, and explore the expressiveness of its features for classification,
- we explore multi-task and transfer learning strategies to combine classification and counting to help in distinguishing healthy from NOD samples.

II. RELATED WORK

To the best of our knowledge, deep learning approaches have not yet been explored in better understanding disease formation and progression of type 1 diabetes at the level of β-cell granules.

[1] Graz University of Technology, Austria.
[2] Silicon Austria Labs, TU-Graz SAL DES Lab, Austria.
[3] Core Facility Ultrastructure Analysis, Graz, Austria.
[4] Medical University of Graz, Austria.
[5] Center for Biomarker Research in Medicine GmbH, Graz, Austria.
Corresponding Author: lea.bogensperger@icg.tugraz.at

Interestingly, a similar idea was proposed by Zhang et al. [19], where the authors investigated monkeys with type 2 diabetes with Metabolic syndrome. After using binary segmentation followed by a watershed transform, they were able to obtain granule instances allowing further post-processing on important features regarding the granules appearances such as radius and size of granular core and vesicle.

Regarding type 1 diabetes, extensive research can be found that is focused on predicting development of diabetes by means of machine learning, which was partly also analyzed by a survey conducted recently [16]. Tripathi et al. [14] and Xue et al. [17] aim to predict the development of type 1 diabetes by incorporating risk factors such as blood pressure and blood glucose level by using Random Forests and Support Vector Machines based on statistical measures such as accuracy and precision. On the other hand, Zaitcev et al. [18] train a neural network to model long term average glucose levels from 5-12 weeks of daily measurements whereas Alfian et al. [1] focus on modeling blood glucose levels in the next 30-60 minutes to combat hypoglycemia, a condition with too low blood glucose levels that can become dangerous for affected patients. None of these approaches have been applied to electron micrograph images, which are not applicable in situ in clinical practice.

These feature-based approaches are obviously very relevant in practice to predict the risk of patients developing type 1 diabetes or to model blood glucose levels of patients that already suffer from the disease. Deep learning techniques can be very helpful to enhance prediction and diagnosis of disease development and progression and thus improve living conditions of those affected. Indeed, these approaches can also be used in the process of better understanding the nature of type 1 diabetes and the mechanisms that are triggered at the levels of insulin-producing granules within the β-cells. However, research on this data can hardly be carried out on humans since the tissue cannot be harvested from living humans and donors from deceased patients are often at a stage where the disease has progressed further and no granules are left within the β-cells. Therefore, conducting this research on mice is a well-suited alternative that allows for results to advance the process of better understanding the pathogenesis of type 1 diabetes.

III. METHODOLOGY

A. Classification and Granule Counting

Discrimination between healthy and NOD tomographic β-cell images can be defined as a binary classification problem. We propose to learn an encoder – or feature extractor – which provides a feature representation of the input images onto a latent space that can be used to solve the classification and granule counting tasks (see Figure 2). In the proposed setting, data is scarce due to its acquisition and annotation costs. Therefore, we propose to use ResNet-18 [5] initialized with pre-trained weights on ImageNet [4] as the encoder. Pre-trained models on large datasets have shown to provide good feature representations that allow for a more robust initialization and faster training when fine-tuning on small

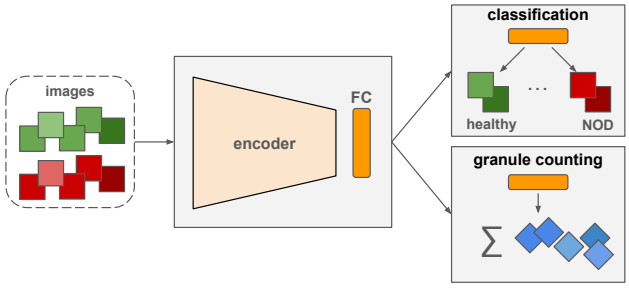


Fig. 2. Proposed architecture to learn classification and/or granule counting.

tasks [8]. Since the output feature space of those pre-trained models are usually of high dimensionality, we propose to include a fully-connected layer between the backbone and the classifiers to enforce a dimensionality reduction. The objective function that is minimized to learn the network parameters is the binary cross-entropy loss

$$\mathscr{L}_{\mathrm{C}}(y,\hat{y}) = -\frac{1}{N}\sum_{n=1}^{N}(y_n \log(\hat{y}_n) + (1-y_n)\log(1-\hat{y}_n)), \quad (1)$$

where y_n is the true label and $\hat{y}_n$ the predicted label for each sample n. Meanwhile, for counting the number of granules present in an input image the problem is framed as a regression task. Therefore the backbone of the feature encoder is appended with a fully-connected layer, where the network's output is the estimated number of granules. In this case the mean squared error is used as the objective function during learning

$$\mathscr{L}_{\mathrm{R}}(y,\hat{y}) = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2, \quad (2)$$

where a strong penalization for higher deviations in the predicted number of granules from the groundtruth is desirable.

B. Multi-task and Transfer Learning

We assume that granule counting can help discriminate between healthy and NOD samples – we show that this assumption holds in Section IV-B. Traditionally, there are two popular strategies in which both tasks can be incorporated into the learning process to benefit each other during training. The first is multi-task learning [12], where both tasks are learned together at the same time on different heads from the shared fully-connected layer. To this end, the loss function is the weighted combination of the binary cross-entropy loss and the mean-squared error

$$\mathscr{L}_{\mathrm{joint}}(y,\hat{y}) = \lambda\mathscr{L}_{\mathrm{C}}(y,\hat{y}) + (1-\lambda)\mathscr{L}_{\mathrm{R}}(y,\hat{y}). \quad (3)$$

where $\lambda \in (0,1)$ is the trade-off between the two tasks. Learning both tasks ensures that the learned, shared feature space contains meaningful representations for both tasks, either helping each other with shared discriminative features, or with regularization over the capacity and importance of the learned features.

Further, as mentioned above, fine-tuning over pre-trained models has been shown to be a useful means of transferring

good mid-level representations [8]. This idea can be extended beyond models trained on large datasets to smaller tasks also via transfer learning [9]. Therefore, we propose that the backbone and fully-connected layer are trained on the regression task to then use the resulting model as an initialization for learning the classification task – each task learned on the corresponding head. The underlying intuition is that features which have been learned during granule counting can be beneficial for the classification process since these tasks are shown to be related by the discriminative capacity of the number of granules for healthy and NOD samples.

C. Evaluation Metrics

The accuracy for binary classification is given by

$$\mathscr{A}(y,\hat{y}) = \frac{1}{N}\sum_{n=1}^{N}\delta(y_n,\hat{y}_n),$$
$$\text{with} \quad \delta(i,j) = \begin{cases} 1 & \text{if} \quad i=j, \\ 0 & \text{else}, \end{cases} \tag{4}$$

where δ is the indicator function. To quantitatively evaluate the potential of the estimated number of granules to classify between healthy and NOD samples, another evaluation criterion is required. We propose a new metric to compute regression accuracy which searches for the optimal threshold θ^* of granules per sample. This threshold separates below and above which samples are predicted to be NOD or healthy, respectively. On training data, it is found by

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\, \mathscr{A}(y,\hat{y}_\theta), \tag{5}$$

where $\hat{y}_\theta$ is the predicted label based on the threshold θ. Then, once fixed, samples from the test set are classified solely based on the number of counted granules, whereupon the accuracy in Equation 4 can be computed.

D. Dataset

Our underlying objective to the presented methodology is to gain novel insights into the mechanisms of β-cell loss via use of neural networks capable of analyzing images containing insulin-producing β-cells. Autophagy, apoptosis, and endoplasmatic reticulum stress are indicators of stress responses in NOD β-cells after immune system attacks, which are expected to have an influence on the produced insulin through a changed appearance of the granules themselves or through the overall number of granules that are available for insulin production. Therefore, we created a dataset to obtain electron micrographs of pancreata of healthy and NOD mice at different ages. Ultrathin sections (70 nm) are stained with platine blue and lead citrate. Non-overlapping areas of pancreatic islets are then pre-selected and analysed by transmission electron microscopy (FEI Tecnai G2) at 120kV. All images were acquired with a magnification of $2,500\times$ and have a resolution of 1024×1024.

The dataset consists of 362 tomographic images from β-cells of healthy C57BL/6J and NOD mice (3/3), which are widely used in type 1 diabetic research [7], [10]. An exemplary image of each group is shown in Figure 1. The slices are non-overlapping and cover several distinct β-cells within each pancreas. The dataset is split into 200 training images and 62 testing images, while ensuring that available data from each individual mouse is either part of the train or the test set.

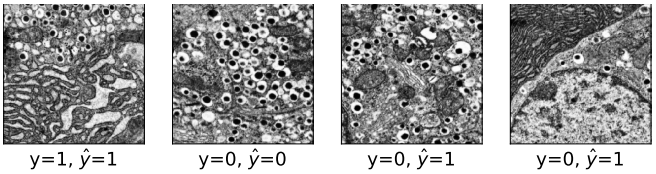


Fig. 3. Examples of correctly classified (left) and misclassified (right) samples from the test set, where y denotes the true class label, and $\hat{y}$ the predicted label. Class labels 0 and 1 denote healthy and NOD, respectively.

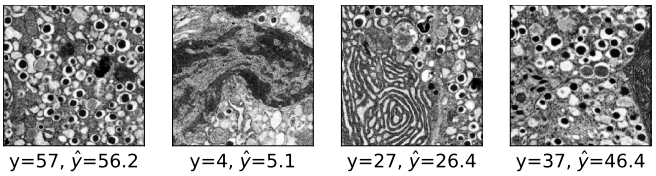


Fig. 4. Examples of granule counting estimation on samples from the test set, where y denotes the groundtruth number of granules, and $\hat{y}$ the predicted number. The rightmost image is the one with the highest counting error on test, probably due to the presence of immature granules and granules from α-cells, which were not included in the training data.

E. Experimental Setup

The chosen encoder backbone is ResNet-18 [5] pre-trained on ImageNet [4]. On top of the encoder, a shared fully-connected layer of size 512×256 is added to reduce the dimensionality of the latent space before the regression and the classification head. By definition of the losses in Section III-A, the first head has 2 outputs representing the binary classification probabilities of healthy and NOD, while the second head has a single output which predicts the number of granules present in the image.

Input images are pre-processed in the same fashion as the encoder backbone was pre-trained. Images are resized to 224×224 followed by standard data augmentation techniques consisting of horizontal and vertical flips. In addition, histogram equalization is applied to all images to compensate for uneven illumination and differences in microscopic settings. During training, we use Adam optimization [6] with an initial learning rate of 1e-4 for classification and 1e-3 for counting regression, and exponential decay rates of $\beta_1=0.9, \beta_2=0.999$.

IV. RESULTS

A. Classification

Classification results on the test set achieved an accuracy of 93.54%. Figure 3 shows examples of correctly and misclassified samples, where class 0 and class 1 indicate healthy and NOD, respectively.

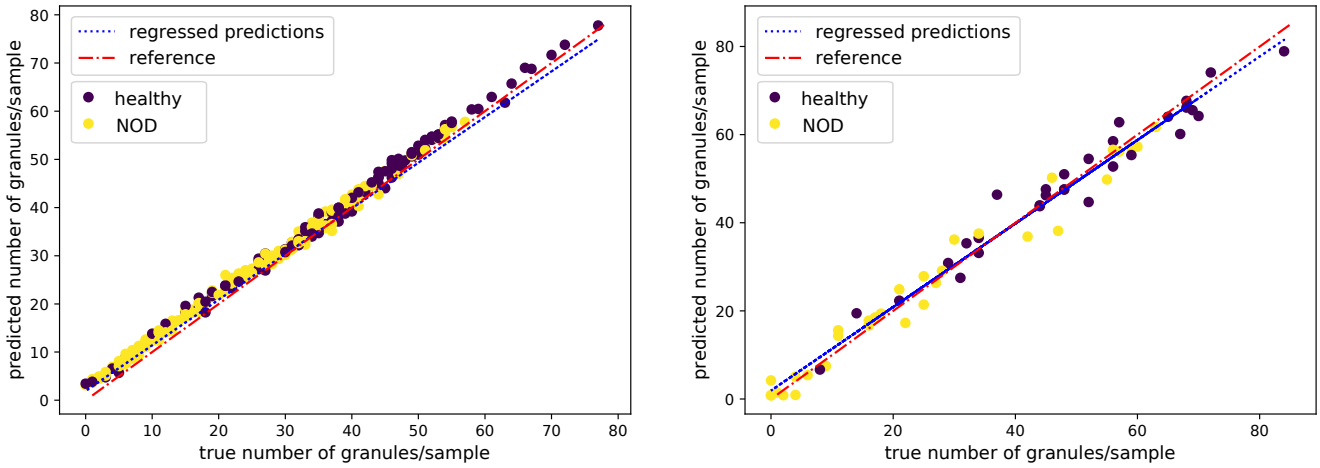


Fig. 5. For both the train (left) and test set (right) in the regression task, the predicted vs. the groundtruth number of granules are displayed with their true class labels. The tendency that healthy samples contain more granules than NOD can be visually observed in both cases.

B. Granule Counting

Training the model on granule counting yields a mean absolute error of 2.78. Examples of granule estimation on the test set are shown in Figure 4. Deviations from the groundtruth number of granules might arise due to the fact that β-cells contain other cellular components with circular structures that may resemble granules and are thus hard to distinguish. The maximum absolute error obtained across all test images is 9.38, shown at the rightmost of Figure 4. This image contains some immature granules and α-cell granules, which the network has not learned to discriminate from β-cell granules due to the limited presence of α-cell granules during training. Furthermore, it has to be noted that it is challenging even for human observers to agree on the exact number of granules, since some imaged granules are out of focus or subject to some microscopic acquisition artefacts. Therefore, we argue that the predicted number of granules is within reason to assume that the model has learned the task properly.

Figure 5 shows the predicted number of granules per sample with respect to the groundtruth. We observe a similar light underestimation on both train and test for images with more granules and a neglectable overestimation for images with lower granule count. Furthermore, each sample is labeled with its true class label, where for both train and test data a relation between a larger/smaller number of granules and a healthy class label 0/NOD class label 1 can be observed.

C. Granule Counting for Classification

As seen in the previous experiment, using the number of estimated granules in a given sample can be an approximate indicator to classify it as either healthy or NOD. Therefore, we apply the optimal threshold search proposed in Section III-C to the granule counting model results on training data. We provide the curve profile of the training data and the corresponding optimal threshold θ^* in Figure 6.

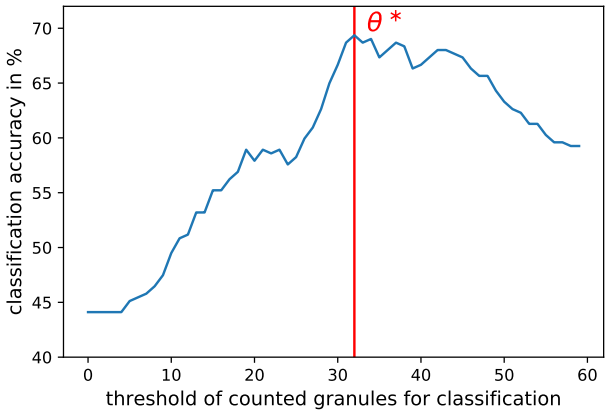


Fig. 6. Evaluation of the classification accuracy on samples, where solely the regression task of counting granules has been trained. Classification accuracy is obtained by thresholding at θ number of granules per slice. Above that threshold value a sample is considered as healthy. Setting the threshold to $\theta^* = 32$ granules/slice yields the maximum classification accuracy of 69.36% within the training data.

Classification results on the training data yield a maximum accuracy of 69.36% ($\theta^* = 32$). With the learned $\theta^* = 32$ applied to the test data, an accuracy of 74.19% is achieved – that is separating solely by granule count. The fact that the accuracy on test is higher than train highlights how the model was not directly trained for the classification task, although it can perform significantly well on it. Doing a Welch's t-test [15] on the estimated number of granules per sample based on the true class label yields a t-statistic of 4.96 with a corresponding p-value of 6.3×10^{-6}, which further emphasizes the strong significant difference in number of granules between healthy and NOD.

D. Joint Granule Counting and Classification

Following the results presented above, granule counting can help distinguish between healthy and NOD images. As

TABLE I

SUMMARY OF THE RESULTS ON THE INFLUENCE OF β-CELL GRANULE COUNTING FOR CLASSIFICATION.

	test accuracy
only classification	93.54%
only counting regression	74.19%
joint ($\lambda = 0.8$)	95.16%
transfer learning (counting $\rightarrow$ classif.)	96.77%

mentioned in Section III-B, we can then exploit this synergy to propose a joint training of both tasks. We simultaneously train both tasks under the same architecture using Equation 3 and with the balancing parameter λ. A grid search over $\lambda \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$ shows that the best result on the train set is obtained at $\lambda = 0.8$. Setting $\lambda = 0$ or $\lambda = 1$ yields the baselines of training only regression or classification, respectively. As expected, classification is given more emphasis since it is the main task at which we evaluate, and we hypothesize that counting granules helps in classification in the sense of a regularizer that learns a better feature representation. This joint multi-task training achieves a classification accuracy of 95.16%, which improves with respect to solely classification or regression, as summarized in Table I.

E. Transfer Learning

Finally, we apply transfer learning from the network trained on granule counting by training classification on top of it. This strategy achieves an accuracy of 96.77%, which surpasses all other results (see Table I). Although the performance of transfer learning is slightly above that of the joint training, we can not exclude that an unexplored λ trade-off exists, which provides a better accuracy. In conclusion, both strategies which exploit the knowledge provided by counting granules when learning to discriminate between healthy and NOD samples outperform the strategies which only rely on a single task.

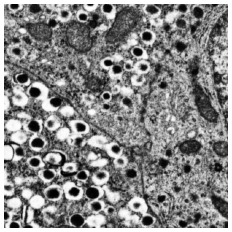
Table II shows an instance where granule counting has a positive impact on the healthy/NOD classification for both joint training and transfer learning. Both healthy samples were predicted as NOD by using only classification. Due to granule counting, which for both samples is rather high, the prediction changes to the correct class when evaluated on the joint or transfer learning models. In the case of joint training, we also provide the prediction for counting.

V. CONCLUSION AND FUTURE WORK

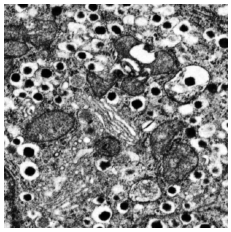
We use a deep learning based approach to achieve high classification accuracy when discriminating between healthy and NOD samples of β-cells. We further show that there is a strong link between the number of granules in a sample and the healthy/NOD class it belongs to. We apply this insight to train a neural network that jointly learns both classification and counting tasks, to improve performance on classification. Further, we explore the concept of transfer learning where a network pre-trained on granule counting can be beneficial

TABLE II

POSITIVE INFLUENCES OF β-CELL GRANULE COUNTING ON CLASSIFICATION FOR JOINT TRAINING (LEFT) AND TRANSFER LEARNING (RIGHT).



GT	only classification	joint
48	-	45.6
0	1	0



GT	only classification	transfer
-	-	-
0	1	0

when fine-tuning it on classification, which also yields an improved accuracy. The multi-tasking and transfer learning approaches allow for a more robust representation by including the additional counting task. This can be especially useful in the medical domain, where often only limited data is available but related additional tasks can be defined in a similar fashion.

The findings strongly support the underlying hypothesis that early onset diabetes leads to a reduction in insulin-producing granules. Nevertheless, classification on its own already delivers strong results. This indicates that there are additional factors other than the number of granules that play a major role in the decision process, such as changed appearances in the granules themselves or the surrounding tissue. Therefore, it is planned to expand this analysis to a larger scale once more data has been acquired, which can be a time-consuming process, but will allow to draw conclusions that are statistically reliable.

VI. ETHICAL APPROVAL

All mouse procedures were in accordance with institutional guidelines and approved by the corresponding authorities (Projects BMWFW-66.010/0142-WF/V/3b/2017 for NOD/ShiLtJ mice, BMWFW-66.010/0160-WF/V/3b/2017 for C57BL/6J mice).

ACKNOWLEDGMENT

This work was supported by the BioTechMed Graz flagship project "MIDAS". Marc Masana acknowledges the support by the "University SAL Labs" initiative of Silicon Austria Labs (SAL).

REFERENCES

[1] G. Alfian, M. Syafrudin, J. Rhee, M. Anshari, M. Mustakim, and I. Fahrurrozi, "Blood glucose prediction model for type 1 diabetes based on extreme gradient boosting," *IOP Conference Series: Materials Science and Engineering*, vol. 803, p. 012012, 2020.

[2] M. A. Atkinson, G. S. Eisenbarth, and A. W. Michels, "Type 1 diabetes," *The Lancet*, vol. 383, no. 9911, pp. 69–82, 2014.

[3] A. Burrack, T. Martinov, and B. Fife, "T cell-mediated beta cell destruction: Autoimmunity and alloimmunity in the context of type 1 diabetes," *Frontiers in Endocrinology*, vol. 8, 2017.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 248–255.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 770–778.

[6] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[7] C. E. Mathews, S. Xue, A. Posgai, Y. L. Lightfoot, X. Li, A. Lin, C. Wasserfall, M. J. Haller, D. Schatz, and M. A. Atkinson, "Acute versus progressive onset of diabetes in nod mice: Potential implications for therapeutic interventions in type 1 diabetes," *Diabetes*, vol. 64, p. 3885–3890, 2015.

[8] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2014.

[9] S. J. Pan and Q. Yang, "A survey on transfer learning," *Transactions on Knowledge and Data Engineering*, vol. 22, pp. 1345–1359, 2010.

[10] J. Pearson, F. Wong, and L. Wen, "The importance of the non obese diabetic (nod) mouse model in autoimmune diabetes." *Journal of autoimmunity*, vol. 66, pp. 76–88, 2016.

[11] B. O. Roep, "There is something about insulin granules," *Diabetes*, vol. 69, no. 12, pp. 2575–2577, 2020.

[12] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[13] H. S. Spijker, R. B. Ravelli, A. M. Mommaas-Kienhuis, A. A. van Apeldoorn, M. A. Engelse, A. Zaldumbide, S. Bonner-Weir, T. J. Rabelink, R. C. Hoeben, H. Clevers *et al.*, "Conversion of mature human β-cells into glucagon-producing α-cells," *Diabetes*, vol. 62, no. 7, pp. 2471–2480, 2013.

[14] G. Tripathi and R. Kumar, "Early prediction of diabetes mellitus using machine learning," in *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 2020, pp. 1009–1014.

[15] B. L. Welch, "The generalisation of student's problems when several different population variances are involved." *Biometrika*, vol. 34 1-2, pp. 28–35, 1947.

[16] A. Z. Woldaregay, E. Årsand, T. Botsis, D. Albers, L. Mamykina, and G. Hartvigsen, "Data-driven blood glucose pattern classification and anomalies detection: machine-learning applications in type 1 diabetes," *Journal of medical Internet research*, vol. 21, no. 5, p. e11030, 2019.

[17] J. Xue, F. Min, and F. Ma, "Research on diabetes prediction method based on machine learning," *Journal of Physics: Conference Series*, vol. 1684, p. 012062, 2020.

[18] A. Zaitcev, M. R. Eissa, Z. Hui, T. Good, J. Elliott, and M. Benaissa, "A deep neural network application for improved prediction of hba1c in type 1 diabetes," *Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 2932–2941, 2020.

[19] X. Zhang, X. Peng, and C. e. a. Han, "A unified deep-learning network to accurately segment insulin granules of different animal models imaged under different electron microscopy methodologies," *Protein Cell*, vol. 10, 2019.

Computed Tomography Reconstruction Using Generative Energy-Based Priors

Martin Zach[1], Erich Kobler[2], and Thomas Pock[1]

***Abstract*— In the past decades, Computed Tomography (CT) has established itself as one of the most important imaging techniques in medicine. Today, the applicability of CT is only limited by the deposited radiation dose, reduction of which manifests in noisy or incomplete measurements. Thus, the need for robust reconstruction algorithms arises. In this work, we learn a parametric regularizer with a global receptive field by maximizing it's likelihood on reference CT data. Due to this unsupervised learning strategy, our trained regularizer truly represents higher-level domain statistics, which we empirically demonstrate by synthesizing CT images. Moreover, this regularizer can easily be applied to different CT reconstruction problems by embedding it in a variational framework, which increases flexibility and interpretability compared to feed-forward learning-based approaches. In addition, the accompanying probabilistic perspective enables experts to explore the full posterior distribution and may quantify uncertainty of the reconstruction approach. We apply the regularizer to limited-angle and few-view CT reconstruction problems, where it outperforms traditional reconstruction algorithms by a large margin.**

I. INTRODUCTION

Throughout the past decades, Computed Tomography (CT) has become an invaluable tool in diagnostic radiology. However, along with its ever-increasing usage have come concerns about the associated risks from ionizing radiation exposure [6]. Approaches that try to remedy this problem include hardware measures such as tube current reduction or modulation (for instance in the form of automatic exposure control [37]), adaptive section collimation [15], or angular under-sampling [11], [10]. Such measures are now standard in clinical CT systems, but require robust reconstruction algorithms.

Classical CT reconstruction algorithms include Filtered Back-Projection (FBP) [8], [18], which has been superseded by more robust iterative algebraic reconstruction techniques [36], [40] in clinical practice. In light of dose reduction, these algorithms may be equipped with prior knowledge to increase reconstruction quality of low-dose scans. Traditional, hand-crafted regularizers, such as Total Variation (TV) [35] and extensions such as Total Generalized Variation (TGV) [5], typically encode regularity assumptions of the reconstruction, such as sparsity of gradients. These hand-crafted regularizers have been used extensively and successfully in reconstruction problems [13], [26], [41], however they do not fully model the a-priori available information. To capture also higher-order image statistics, the idea of learning a regularizer from data emerged [43], [34], [23]. Although these learning-based approaches are now dominant in many fields, such models have classically focused on modeling local statistics and leave much to be desired in modeling global dependencies.

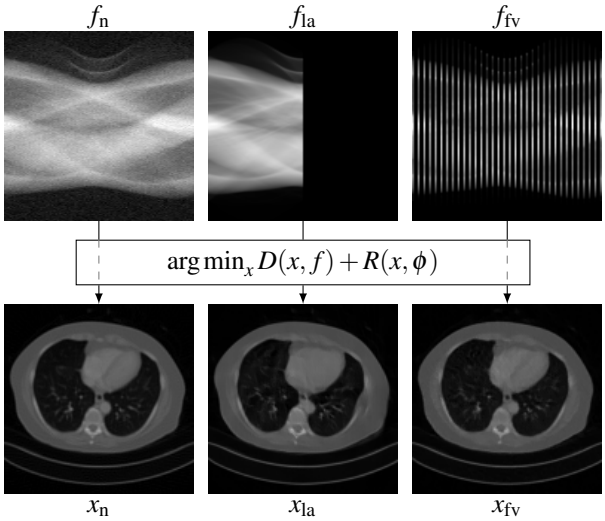


Fig. 1. Our proposed method is able to reconstruct images from noisy, limited-angle and few-view measurements (denoted by the subscripts $\mathrm{n}, \mathrm{la}, \mathrm{fv}$) satisfactorily.

From a statistical point of view, any regularizer R induces a Gibbs-Boltzmann distribution

$$p_R(x) = \frac{\exp(-R(x))}{\int_{\mathscr{X}} \exp(-R(\xi))\ \mathrm{d}\xi}, \tag{1}$$

where $\mathscr{X}$ is the space of all possible images. Ideally, samples $x \sim p_R$ should be indistinguishable from samples from the underlying reference distribution, which is hardly possible for hand-crafted regularizers.

In this work, we propose a novel generatively trained regularizer utilizing a global receptive field that yields high-quality reconstructions even in case of strong noise or heavily undersampled measurements. In Fig. 1, we show how our model is able to satisfactorily reconstruct CT images from noisy (i.e. low tube current) and incomplete (i.e. limited-angle or few-view) data without observable artifacts. In fact, using this regularizer we can synthesize naturally appearing CT images *without any* data (see Fig. 4). In contrast to feed-forward formulations [4], [12], we cast the reconstruction as a variational problem. This helps interpretability of the trained regularizer by means of analyzing its induced distribution as

[1]Institute of Computer Graphics and Vision, Graz University of Technology, 8010 Graz, Austria {martin.zach,pock}@icg.tugraz.at
[2]Institute of Computer Graphics, Johannes Kepler University Linz, 4040 Linz, Austria erich.kobler@jku.at

well as the posterior distribution of any type of reconstruction problem. We apply a trained model to limited-angle and few-view reconstruction problems, and compare our approach quantitatively and qualitatively with traditional reconstruction algorithms. In addition, we perform experiments which leverage the probabilistic nature of our approach, such as prior and posterior sampling.

To summarize, we

- define a novel network architecture capable of synthesizing natural CT images without measurement data,
- demonstrate that our regularizer outperforms classical algorithms in typical reconstruction problems, and
- show that our probabilistic approach allows to compute the pixel-wise posterior-variance, which in turn is related to uncertainty quantification.

II. Related Work

A. Learning-based CT Reconstruction

In recent years, there has been a strong shift from hand-crafted regularizers towards data-driven reconstruction schemes. The learning-based methods can be applied in the sinogram domain [4], [19], such that the final image can be reconstructed using traditional reconstruction algorithms. Alternatively, a preliminary reconstruction may be computed using the (noisy and possibly incomplete) sinogram, which can subsequently be enhanced by a trained convolutional neural network (CNN) [12]. An alternative learning-based reconstruction approach is to learn a direct mapping from the data domain to the image domain [42]. However, this requires to learn a wealth of parameters solely to compute an approximate inverse of the forward acquisition operator. Another recently popularized approach is to learn an unrolled iterative reconstruction algorithm [20], [2], [24]. Whilst the results look promising, we point out that such approaches typically assume a particular acquisition setup and, at inference time, can only be applied in settings that are very similar to the training setting.

B. Generative Models as Regularizers in Medical Imaging

Energy-based models (EBMs) have a long history in the field of image processing [25]. However, only recently some works [17], [29] have explored their generative capabilities, rivaling the performance of Generative Adversarial Networks (GANs). While GANs have been used as an implicit prior for reconstruction problems in medical imaging (e.g. [1]), to the best of our knowledge, using EBMs capable of synthesizing natural images at full-scale as regularizers in medical imaging is still largely unexplored.

III. Methodology

In this work, we represent CT images of size $n = n_w \times n_h$ pixels as vectors $x \in \mathbb{R}^n$. The subsequent analysis easily generalizes to image data in any dimensions. Acquiring n_θ projections with n_d detector elements, the post-log sinogram $f \in \mathbb{R}^m$ of size $m = n_\theta \times n_d$ is given by

$$f = Ax + \eta, \tag{2}$$

where $A\colon \mathbb{R}^n \to \mathbb{R}^m$ is the acquisition operator, and $\eta \in \mathbb{R}^m$ represents the additive measurement noise, summarizing photon statistics, thermal noise in the measurement channels, and pre-processing steps. The linear acquisition operator A is defined by the geometry of the measurement setup, and throughout this work we assume that both A and η can be characterized up to reasonable precision.

A. Bayesian Modeling

To account for measurement uncertainties and missing data in the observations f, we adopt a rigorous statistical interpretation of (2). Bayes' Theorem relates the posterior probability $p(x \mid f)$ to the data-likelihood $p(f \mid x)$ and the prior $p(x)$ by

$$p(x \mid f) \propto p(f \mid x)p(x). \tag{3}$$

Here, $p(x \mid f)$ quantifies the belief in a solution x given a datum f. In the negative log-domain, (3) is transformed to

$$E(x,f) \coloneqq D(x,f) + R(x), \tag{4}$$

where we identify the *data-fidelity* term $D\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^+$ modeling the negative log-likelihood $-\log p(f \mid x)$, and the *regularizer* $R\colon \mathbb{R}^n \to \mathbb{R}$ modeling the negative log-prior $-\log p(x)$. The *energy* $E\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ assigns a scalar $E(x,f)$ to any (x,f)-pair, and in the sense of (3) is interpreted as the negative log-posterior $-\log p(x \mid f)$.

Typically, D makes use of the forward operator A to quantify the agreement between the reconstruction of x and the measured data f. R may for instance represent the TV semi-norm [35], which is well known to favor piece-wise constant solutions. For the sake of simplicity, we assume η to be Gaussian, and consequently set $D(x,f) = \frac{1}{2\sigma^2}\|Ax - f\|^2$, where σ^2 denotes the variance of η. We discuss the choice of R in the next section.

B. Parameter Identification

Although many hand-crafted choices for R exist, such as TGV [5] or wavelet-based approaches [16], it is generally agreed upon that modeling higher order image statistics should be based on learning [43]. In contrast to the widely adopted feed-forward approaches, in this work we retain the variational structure to allow statistical interpretation. To account for the parameters, we extend (4) to

$$E(x,f,\phi) \coloneqq D(x,f) + R(x,\phi), \tag{5}$$

where $R\colon \mathbb{R}^n \times \Phi \to \mathbb{R}$ is parametrized by ϕ in the set of feasible parameters Φ. We illustrate our particular choice of R (for two-dimensional input images) in Fig. 2 and emphasize that the input image is reduced to a scalar only by means of (strided) convolutions. Here, ϕ summarizes the convolution kernels and biases, and Φ reduces to $\mathbb{R}^{n_p}$, where n_p is the total number of parameters.

The Bayesian separation of of data-likelihood and prior allows us to train our regularizer *generatively* without any measurement data as follows. We denote by p_ϕ the Gibbs-Boltzmann distribution of $R(\cdot,\phi)$ in the sense of (1), to emphasize the dependence on the parameters. Assuming

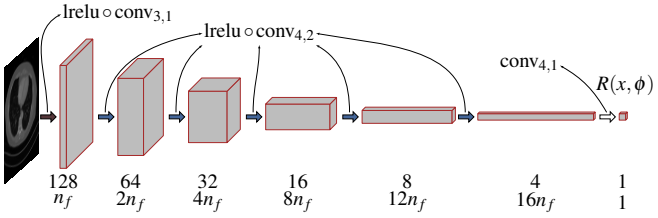


Fig. 2. Our proposed architecture follows a typical encoder structure. The subscripts specify filter size and stride and the annotations show the spatial resolution of the feature maps and the number of features.

access to a distribution p_x of reference CT images, we identify the optimal parameters ϕ^* by minimize the negative log-likelihood

$$\phi^* \in \arg\min_{\phi\in\Phi}\{\Gamma(\phi) := \mathbb{E}_{x\sim p_x}[-\log p_\phi(x)]\}. \tag{6}$$

In the seminal work of [21], it is shown that the gradient of (6) with respect to the parameters ϕ is given by

$$\nabla_1\Gamma(\phi) = \mathbb{E}_{x^+\sim p_x}[\nabla_2 R(x^+,\phi)] - \mathbb{E}_{x^-\sim p_\phi}[\nabla_2 R(x^-,\phi)], \tag{7}$$

where ∇_l denotes the gradient w.r.t. the l-th argument. We discuss the estimation of the expectations in both terms extensively in Sec. III-C.

We highlight that (6) does not require any (x, f)-pairs. That is, for training we do not require access to measurement data but only to (the usually much more ubiquitous) reference images. Moreover, a trained regularization model serves as a drop-in replacement for hand-crafted regularizers for any reconstruction problem by adapting the data-fidelity D to account for a particular forward operator A and noise statistics.

C. Model Sampling

While the first term in (7) is easily approximated given any dataset, the second term requires sampling the induced model distribution, which is known to be hard in high dimensions [7]. For any reasonably sized image $x \in \mathbb{R}^n$ computing the partition function is infeasible, hence the distribution has to be approximated using Markov Chain Monte Carlo (MCMC) techniques. In this work, we utilize the unadjusted Langevin algorithm (ULA) [32], [31], [33], which makes use of the gradient of the underlying probability density function to improve mixing times of the Markov chains. The ULA algorithm read as

$$x^k \sim \mathcal{N}(x^{k-1} + \frac{\varepsilon}{2}\nabla_1 \log p_\phi(x^{k-1}), \beta\varepsilon \mathrm{Id}_n),\ k = 1,\dots,K, \tag{8}$$

where $\mathcal{N}(\mu,\Sigma)$ denotes the normal distribution on $\mathbb{R}^n$ with mean μ and covariance Σ. $\beta, \varepsilon \in \mathbb{R}^+$ are appropriately chosen scaling parameters, and K denotes the total number of steps. To aid the convergence of the Markov chains, we further follow the idea of persistent chains [38] and use a buffer in which the states of the chains persist throughout parameter updates.

D. Experimental Setup

For all the following experiments, we set $n_f = 48$, resulting in $n_p = 12\,179\,905$ and set the ReLU leak coefficient to 0.05.

Algorithm 1: Maximum Likelihood training of an EBM. $\mathcal{U}(\mathcal{X})$ denotes the uniform distribution on $\mathcal{X}$ and each r denotes an independent sample from $\mathcal{U}([0,1])$.

Input : p_x, σ_{data}, n_{buffer}, p_{re}, K, ϕ, n_e, ε, β
Output: ϕ approximately minimizing (6)

$\mathcal{B} \leftarrow \{u_1,\dots,u_{n_{\text{buffer}}}\}$, $u_i \sim \mathcal{U}([0,1]^n)$
for $t = 1,\dots,n_e$ **do**
  $x^+ \sim (p_x * \mathcal{N}(0, \sigma^2_{\text{data}}\mathrm{Id}_n)), x^0 \sim \mathcal{B}$
  Generate x^- with (8) using x^0, ε, K, β
  if $r > p_{re}$ **then** $x_{\text{refill}} = x^-$
  else
    if $r > 0.5$ **then** $x_{\text{refill}} = x^+$
    else $x_{\text{refill}} = u \sim \mathcal{U}([0,1]^n)$
  end
  $\mathcal{B} \leftarrow \mathcal{B} \setminus \{x^0\} \cup \{x_{\text{refill}}\}$
  $\phi \leftarrow \text{Adam}(\nabla_2 R(x^+,\phi) - \nabla_2 R(x^-,\phi))$
end

We trained the regularizer on the Low Dose CT Image and Projection dataset [27], where the images were downsampled to 128×128. We optimized (7) using Adam [22] with a learning rate of 5×10^{-4} and set the first and second order momentum variables to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. To stabilize training, we convolved p_x with $\mathcal{N}(0, \sigma^2_{\text{data}}\mathrm{Id}_n)$, where $\sigma_{\text{data}} = 1.5 \times 10^{-2}$. We used a batch size of 25 and a replay buffer holding 8000 images with reinitialization chance of $p_{\text{re}} = 1\,\%$. Samples in the buffer were reinitialized with an equal chance of uniform noise or samples from the data distribution. To sample p_ϕ, we ran (8) with $K = 500$, using $\varepsilon = 1$ and $\beta = 7.5 \times 10^{-3}$.[1] We summarize the training algorithm in Alg. 1.

For the reconstruction problems, we used accelerated proximal gradient descent [28], as summarized in Alg. 2 with $J = 1 \times 10^3$, $\gamma_1 = 0.5$, $\gamma_2 = 1.5^{-1}$. We solve the proximal operator $\text{prox}\colon \mathbb{R}^n \to \mathbb{R}^n$, which for $H\colon \mathbb{R}^n \to \mathbb{R}$ and $\tau \in \mathbb{R}^+$ is defined as

$$\text{prox}_{\tau H}(y) = \arg\min_x \tau H(x) + \frac{1}{2}\|x-y\|_2^2 \tag{9}$$

using 10 iterations of the conjugate gradient method. In what follows, the forward operator A assumes a parallel-beam geometry with $n_d = 362$ detectors of size 1 pixel and is discretized using the ASTRA toolbox [39]. Unless stated otherwise, η is 0.1 % Gaussian noise.

IV. Results

A. Induced Prior Distribution

For most hand-crafted regularizers, there typically exists a geometrical interpretation. For instance, it is well known that TV is related to the perimeter of the level sets of an image [9]. Hence, the influence on the reconstruction is fairly easily understood. Our regularizer can hardly be interpreted in such

[1] Similar to [30], we reparametrize the regularizer as $\frac{R}{T}$ for a-priori chosen T, such that $\varepsilon = 1$ in (8).

Algorithm 2: Accelerated proximal gradient descent with Lipschitz-backtracking.

Input : initial α, f, x^0, ϕ, J, $\gamma_1 \in (0,1)$, $\gamma_2 \in (0,1)$
Output : x^{J+1} approximately minimizing (5)

$x^1 = x^0$
for $t = 1, \dots, J$ **do**
  $\bar{x} = x^t + \frac{t}{t+3}(x^t - x^{t-1})$
  $g = \nabla_1 R(\bar{x}, \phi)$
  for *ever* **do**
    $x^{t+1} = \operatorname{prox}_{\alpha D(\cdot, f)}(\bar{x} - \alpha g)$
    $Q = R(\bar{x}, \phi) + \langle g, x^{t+1} - \bar{x} \rangle + \frac{1}{2\alpha} \left\| x^{t+1} - \bar{x} \right\|_2^2$
    if $R(x^{t+1}, \phi) \leq Q$ **then**
      $\alpha \leftarrow \alpha / \gamma_1$
      break
    else $\alpha \leftarrow \gamma_2 \alpha$
  end
end

a way, however the energy-perspective allows for a statistical analysis by means of the Gibbs-Boltzmann distribution p_ϕ.

One of the main characteristics of any distribution are the points which locally maximize the density (modes). By (1) it is easily seen that the modes of p_ϕ conincide with local minima of $R(\cdot, \phi)$. However, modes may occur as spikes in regions of generally low mass, and thus samples may represent the underlying distribution more accurately. Therefore, we inspect our regularizer by computing modes as well as samples.

We find $x \sim p_\phi$ using Langevin sampling (8) with $K = 40\,000$ steps, and find $\arg\min_x R(x, \phi)$ with Alg. 2 using $D(x, f) = 0$. In both cases, we set $x_0 \sim \mathcal{U}([0,1]^n)$. We show the trajectories of x^t during minimzation of $R(\cdot, \phi)$ and samples $x \sim p_\phi$ in Fig. 3.

The results indicate that our model is able to synthesize natural CT images without any measurement data. This is in stark contrast to other priors typically used in medical imaging (see e.g. [1, Fig. 1] for samples drawn from hand-crafted priors).

B. Limited-Angle and Few-View Reconstruction

In this section, we shift our focus towards CT reconstruction problems, where we first treat the reconstruction problem as a deterministic mapping in the maximum a-posteriori (MAP) sense. Specifically, we denote by $x^* \colon \mathbb{R}^m \to \mathbb{R}^n$ the model-optimal reconstruction identified by the mapping

$$x^*(f) \in \arg\min_x \{D(x, f) + R(x, \phi)\}. \tag{10}$$

Further, let $p_{\hat{x}}$ denote a distribution on $\mathbb{R}^n \times \mathbb{R}^m$ of (problem-dependent) (f, x)-pairs of a (noisy and incomplete) datum f and the corresponding reference image x.

To illustrate the capabilities of our trained regularizer, we first consider a limited-angle reconstruction problem. Specifically, we reconstruct an image from $n_\theta = 270$ projections uniformly spaced over the quarter-circle $\theta \in [0, \frac{\pi}{2}]$. We show qualitative results in Fig. 4 (top), where the FBP reconstruction exhibits smearing artifacts that are characteristic of limited-angle CT. Simultaneous Algebraic Reconstruction Technique (SART) [3] and additional TV regularization help remedy this problem somewhat, however the reconstruction is not satisfactory. We observe unnatural disconnected contours in the reconstruction, especially around the thorax. On the contrary, our model is capable of reconstructing a natural looking image with realistic anatomy and high level of detail. We show $\mathbb{E}_{(f,x)\sim p_{\hat{x}}}[\mathrm{PSNR}(x^*(f), x)]$ in Tab. I. The results are in accordance with the qualitative analysis, with our model improving the TV reconstruction by over 4.5 dB.

TABLE I
$\mathbb{E}_{(f,x)\sim p_{\hat{x}}}[\mathrm{PSNR}(x^*(f), x)]$ FOR LIMITED-ANGLE ($\theta \in [0, \frac{\pi}{2}]$) AND FEW-VIEW ($n_\theta \in \{100, 50, 30, 20\}$) RECONSTRUCTION.

		FBP	SART	TV	Ours
limited-angle	$\theta \in [0, \frac{\pi}{2}]$	19.05	27.72	29.67	**34.21**
few-view	$n_\theta = 100$	37.15	43.86	46.77	**49.47**
	$n_\theta = 50$	33.12	37.05	40.21	**45.06**
	$n_\theta = 30$	28.78	33.04	35.33	**41.65**
	$n_\theta = 20$	25.24	30.55	31.77	**38.48**

In contrast to limited-angle CT, in few-view CT data are acquired over the full half-circle $\theta \in [0, \pi]$. However, on this half-circle only $n_\theta \ll$ projections are sparsely acquired. In traditional reconstruction algorithms, the sparse data manifests itself as streaking artifacts around sharp contours, where subsequent projections do not properly cancel each other. Such artifacts can clearly be seen in the FBP reconstruction in Fig. 4 (bottom), where we show the results for a $n_\theta = 20$ few-view reconstruction problem. TV regularization yields a sharp and largely artifact-free image at the cost of losing almost all details. Our method can reconstruct the image satisfactorily, where artifacts are removed whilst retaining small details. Tab. I shows quantitative results, with our approach consistently beating the reference methods for all $n_\theta \in \{100, 50, 30, 20\}$ by a large margin.

C. Posterior Analysis

Instead of treating R as a point estimator in the maximum a-posteriori sense (10), the Bayesian formulation allows to explore the full posterior distribution of any given reconstruction problem. This is especially useful in the medical domain, where interpretability is of utmost importance. To this end, we perform Langevin sampling of the posterior distribution (3) with the same parameters as in training. We show some illustrative examples for limited-angle and few-view CT in Fig. 5. The figure shows samples $\xi \sim p(x \mid f, \phi) = p_\phi(x) p(f \mid x)$ from the posterior distribution associated with Eq. (5) as well as it's expectation and variance. For the limited-angle reconstruction, we observe large variance around regions of high ambiguity, where there exist no projections to define contours. Similarly, for the few-view problem, there is high variance around small structures such as the vertebrae or blood vessels in the lung. For both problems, the approximated expected value over the posterior also yields a visually

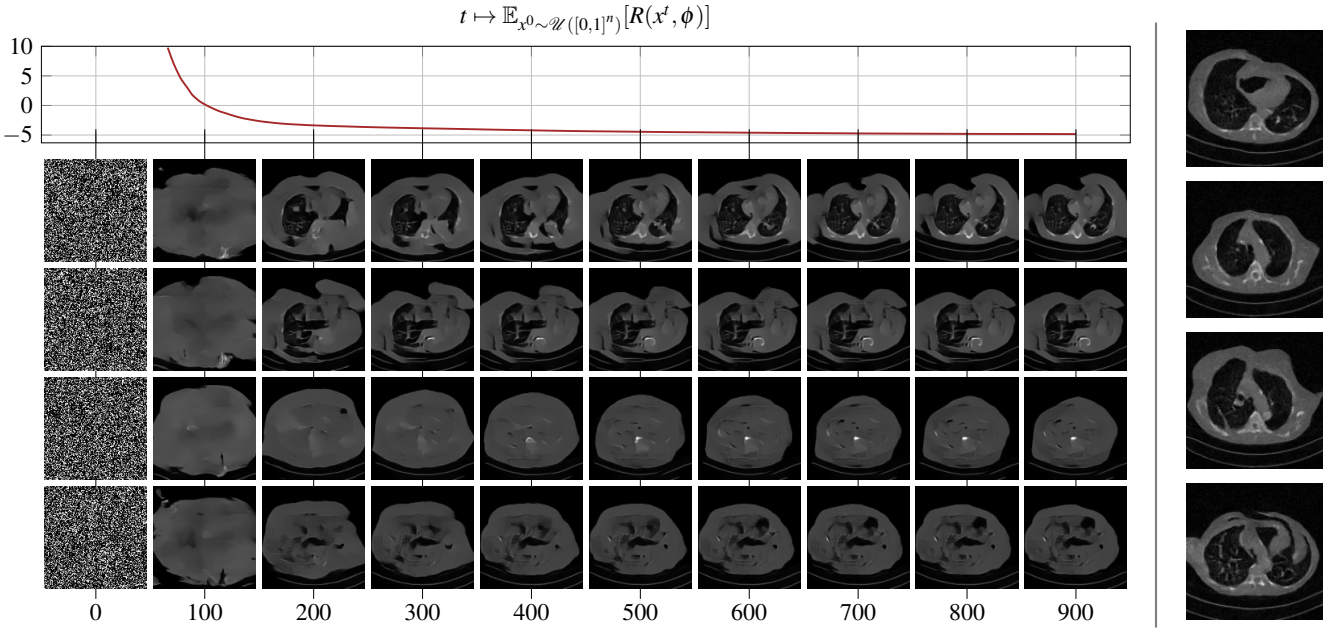


Fig. 3. Trajectories of the images from uniform noise to $\arg\min_x R(x, \phi)$ along with the corresponding $R(x^t, \phi)$ (left) and samples $x \sim p_\phi$ from the Langevin process (8) after $K = 40000$ steps (right).

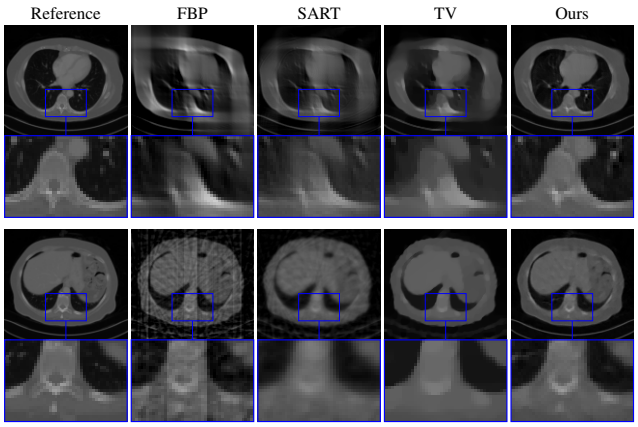


Fig. 4. Comparison between FBP, SART, TV, and our method for limited-angle ($\theta \in [0, \frac{\pi}{2}]$, top) and few-view ($n_\theta = 20$, bottom) CT reconstruction. Our model is able to faithfully reconstruct the image, whereas the other methods are not able to fully remove the smearing and streaking artifacts.

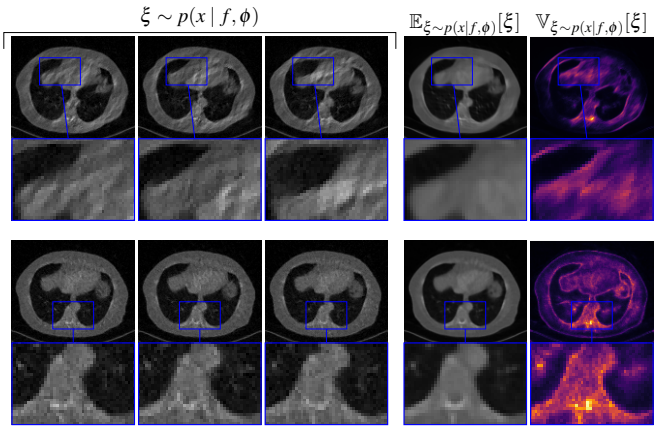


Fig. 5. Sampling the posterior of a limited-angle ($\theta \in [0, \frac{\pi}{2}]$, top) and few-view ($n_\theta = 30$, bottom) CT reconstruction problem: The three images on the left show different samples during the sampling process, the two images on the right show the expected value and variance of the posterior distribution respectively.

appealing, although somewhat over-smoothed, reconstruction.

D. Out-of-Distribution Application

1) Uncertainty Quantification Through Posterior Variance Analysis: To study how the variance relates to uncertainty, we perform the following experiment: We introduce unnatural (read: not present in the training data) structures into the image by overlaying the "cameraman" image and an example of the "grid" texture from the Describable Textures Dataset [14] on a reference scan. Subsequently, we approximate the variance of a few-view reconstruction problem using $n_\theta = 20$ views by Langevin sampling.

We show the expected value and variance over the posterior for the clean and corrupted scans in Fig. 6. Although the bulk of the cameraman shows low variance (and indeed the reconstruction looks natural in these regions), we observe high variance in unnatural regions, such as the artificially introduced corners and the tripod. Similarly, compared to the reference scan the grid overlay leads to high variance in the posterior.

In general, we believe that high posterior variance is related to model uncertainty. To be more specific, we expect high variance if the measurement data suggests structures that are not consistent with the training data. This could potentially aid in detecting pathologies in images.

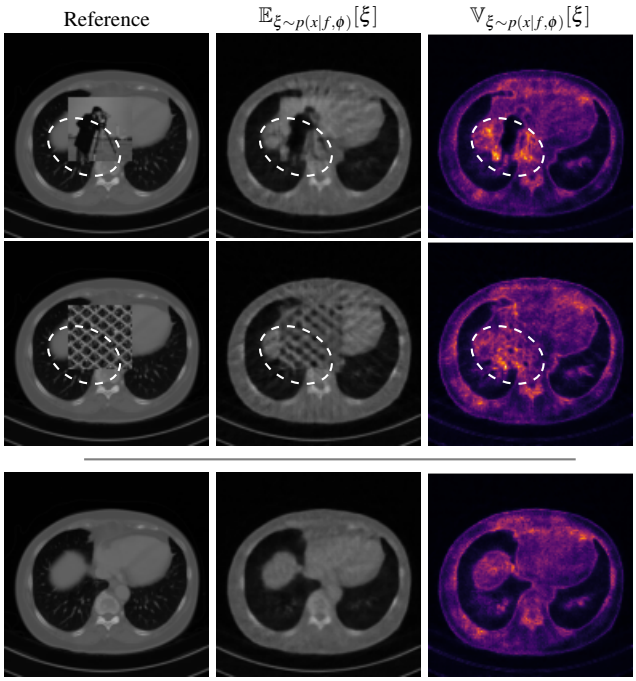


Fig. 6. Comparison of the posterior distribution of a corrupted (top two) versus clean (bottom) scan: The high variance around the corrupted regions (highlighted) relates to high model uncertainty.

2) Generalization: In Sec. IV-A we have shown how samples $x \sim p_\phi$ resemble data drawn from p_x — that is, R encodes a prior in the frequentist sense. With this, a natural question is if our proposed regularizer can be applied to reconstruction problems where the underlying distribution deviates far from p_x. To study this, we propose the following experiment: We let

$$x_\kappa = \mathrm{rot}_\kappa(x) + \eta, \tag{11}$$

where $\mathrm{rot}_\kappa \colon \mathbb{R}^n \to \mathbb{R}^n$ is the bi-linear rotation operator of angle κ and η is 10 % Gaussian noise and find

$$x^* \in \arg\min_x \frac{1}{2\sigma^2} \|x - x_\kappa\|^2 + R(x,\phi). \tag{12}$$

The results in Fig. 7 show that performance quickly deteriorates with increasing κ. This is in line with our expectations, since our regularizer models global characteristics of the reconstruction which are not rotation invariant.

V. Conclusion

In this work, we designed a parametrized regularizer utilizing a global receptive field, which we trained on full-scale CT images by maximizing their likelihood. The induced Gibbs-Boltzmann distribution of the trained regularizer strongly resembles the data distribution — that is, our model is capable of synthesizing natural CT images *without* any data. The maximum likelihood framework does not assume any particular forward acquisition operator or noise statistics, and the trained regularizer can be applied to any reconstruction problem. In limited-angle and few-view reconstruction problems, we observed significantly improved

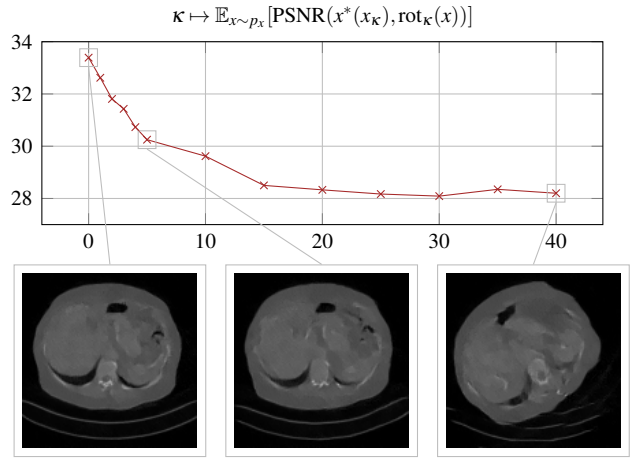


Fig. 7. Performance of the regularizer on out-of-distribution data: For denoising rotated images, the PSNR quickly decays even for small rotations.

quantitative and qualitative performance when compared to classical reconstruction algorithms. Further, we were able to relate the variance in the posterior with unnatural structures in the underlying image, as is the case for certain pathologies.

In summary, we believe that learning energy-based models capable of truly capturing the underlying distribution is a very promising direction for future research. Such models yield natural reconstructions with severely undersampled and noisy data, where data consistency can be enforced with arbitrary data terms. We also want to emphasize that training requires only reconstructed images, which are typically much more ubiquitous than image-data pairs. Future work includes the extension to higher resolutions used in clinical practice today, and tackling the problem of scale- and rotation-invariance. Further, a rigorous mathematical analysis in the context of inverse problems, stability w.r.t. training and measurement data would improve the applicability in clinical practice.

References

[1] J. Adler and O. Öktem, "Deep bayesian inversion," *arXiv preprint arXiv:1811.05910*, 2018.

[2] ——, "Learned primal-dual reconstruction," *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, pp. 1322–1332, 2018.

[3] A. Andersen and A. Kak, "Simultaneous algebraic reconstruction technique (sart): A superior implementation of the art algorithm," *Ultrasonic Imaging*, vol. 6, no. 1, pp. 81–94, 1984.

[4] R. Anirudh, H. Kim, J. J. Thiagarajan, K. A. Mohan, K. Champley, and T. Bremer, "Lose the views: Limited angle ct reconstruction via implicit sinogram completion," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6343–6352.

[5] K. Bredies, K. Kunisch, and T. Pock, "Total generalized variation," *SIAM Journal on Imaging Sciences*, vol. 3, no. 3, pp. 492–526, Jan. 2010.

[6] D. J. Brenner and E. J. Hall, "Computed tomography — an increasing source of radiation exposure," *New England Journal of Medicine*, vol. 357, no. 22, pp. 2277–2284, 11 2007.

[7] S. Brooks, A. Gelman, J. Galin, and X.-L. Menng, *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 2011.

[8] T. M. Buzug, *Computed Tomography: From Photon Statistics to Modern Cone-Beam CT*. Springer Berlin Heidelberg, 2008.

[9] A. Chambolle, V. Caselles, D. Cremers, M. Novaga, and T. Pock, *An Introduction to Total Variation for Image Analysis*. De Gruyter, 2010, pp. 263–340.

[10] B. Chen, E. Kobler, M. J. Muckley, A. D. Sodickson, T. O'Donnell, T. Flohr, B. Schmidt, D. K. Sodickson, and R. Otazo, "SparseCT: System concept and design of multislit collimators," *Medical Physics*, vol. 46, no. 6, pp. 2589–2599, 5 2019.

[11] G.-H. Chen, J. Tang, and S. Leng, "Prior image constrained compressed sensing (PICCS): A method to accurately reconstruct dynamic CT images from highly undersampled projection data sets," *Medical Physics*, vol. 35, no. 2, pp. 660–663, Jan. 2008.

[12] H. Chen, Y. Zhang, M. K. Kalra, F. Lin, Y. Chen, P. Liao, J. Zhou, and G. Wang, "Low-dose ct with a residual encoder-decoder convolutional neural network," *IEEE Transactions on Medical Imaging*, vol. 36, no. 12, pp. 2524–2535, 2017.

[13] Z. Chen, X. Jin, L. Li, and G. Wang, "A limited-angle CT reconstruction method based on anisotropic TV minimization," *Physics in Medicine and Biology*, vol. 58, no. 7, pp. 2119–2141, Mar. 2013.

[14] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi, "Describing textures in the wild," in *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[15] P. D. Deak, O. Langner, M. Lell, and W. A. Kalender, "Effects of adaptive section collimation on patient radiation dose in multisection spiral CT," *Radiology*, vol. 252, no. 1, pp. 140–147, July 2009.

[16] B. Dong, J. Li, and Z. Shen, "X-ray CT image reconstruction via wavelet frame based regularization and radon domain inpainting," *Journal of Scientific Computing*, vol. 54, no. 2-3, pp. 333–349, Feb. 2012.

[17] Y. Du and I. Mordatch, "Implicit generation and modeling with energy based models," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.

[18] L. A. Feldkamp, L. C. Davis, and J. W. Kress, "Practical cone-beam algorithm," *Journal of the Optical Society of America A*, vol. 1, no. 6, pp. 612–619, 6 1984.

[19] M. U. Ghani and W. C. Karl, "Cnn based sinogram denoising for low-dose ct," in *Imaging and Applied Optics*. Optical Society of America, 2018, p. MM2D.5.

[20] K. Hammernik, T. Klatzer, E. Kobler, M. P. Recht, D. K. Sodickson, T. Pock, and F. Knoll, "Learning a variational network for reconstruction of accelerated MRI data," *Magnetic Resonance in Medicine*, vol. 79, no. 6, pp. 3055–3071, Nov. 2017.

[21] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 8 2002.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[23] E. Kobler, A. Effland, K. Kunisch, and T. Pock, "Total deep variation for linear inverse problems," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[24] E. Kobler, A. Effland, T. Pock, B. Chen, and D. Sodickson, "Total deep variation for sparsect reconstruction," in *6th International Conference on Image Formation in X-Ray Computed Tomography*, 2020, 6th International Conference on Image Formation in X-Ray Computed Tomography ; Conference date: 03-08-2020 Through 07-08-2020.

[25] Y. LeCun, S. Chopra, R. Hadsell, F. J. Huang, and et al., "A tutorial on energy-based learning," in *Predicting Structured Data*. MIT Press, 2006.

[26] Y. Liu, Z. Liang, J. Ma, H. Lu, K. Wang, H. Zhang, and W. Moore, "Total variation-stokes strategy for sparse-view x-ray CT image reconstruction," *IEEE Transactions on Medical Imaging*, vol. 33, no. 3, pp. 749–763, Mar. 2014.

[27] T. R. Moen, B. Chen, D. R. Holmes III, X. Duan, Z. Yu, L. Yu, S. Leng, J. G. Fletcher, and C. H. McCollough, "Low-dose ct image and projection dataset," *Medical Physics*, vol. 48, no. 2, pp. 902–911, 2021.

[28] Y. E. Nesterov, "A method for solving the convex programming problem with convergence rate $o(1/k^2)$," *Doklady Akademii Nauk SSSR*, vol. 269, pp. 543–547, 1983.

[29] E. Nijkamp, M. Hill, T. Han, S.-C. Zhu, and Y. N. Wu, "On the anatomy of mcmc-based maximum likelihood learning of energy-based models," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, pp. 5272–5280, Apr. 2020.

[30] E. Nijkamp, M. Hill, S.-C. Zhu, and Y. N. Wu, "Learning non-convergent non-persistent short-run mcmc toward energy-based model," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.

[31] G. O. Roberts and J. S. Rosenthal, "Optimal scaling of discrete approximations to langevin diffusions," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 60, no. 1, pp. 255–268, 2 1998.

[32] G. O. Roberts and R. L. Tweedie, "Exponential convergence of langevin distributions and their discrete approximations," *Bernoulli*, vol. 2, no. 4, pp. 341 – 363, 1996.

[33] P. J. Rossky, J. D. Doll, and H. L. Friedman, "Brownian dynamics as smart monte carlo simulation," *The Journal of Chemical Physics*, vol. 69, no. 10, pp. 4628–4633, 1978.

[34] S. Roth and M. J. Black, "Fields of experts: A framework for learning image priors," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2005, pp. 860–867 vol. 2.

[35] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1, pp. 259–268, 1992.

[36] Y. Saad and H. A. van der Vorst, "Iterative solution of linear systems in the 20th century," *Journal of Computational and Applied Mathematics*, vol. 123, no. 1, pp. 1–33, 2000, numerical Analysis 2000. Vol. III: Linear Algebra.

[37] M. Söderberg and M. Gunnarsson, "Automatic exposure control in computed tomography – an evaluation of systems from different manufacturers," *Acta Radiologica*, vol. 51, no. 6, pp. 625–634, July 2010.

[38] T. Tieleman, "Training restricted boltzmann machines using approximations to the likelihood gradient," in *International Conference on Machine Learning*. New York, NY, USA: Association for Computing Machinery, 2008, p. 1064–1071.

[39] W. van Aarle, W. J. Palenstijn, J. D. Beenhouwer, T. Altantzis, S. Bals, K. J. Batenburg, and J. Sijbers, "The ASTRA toolbox: A platform for advanced algorithm development in electron tomography," *Ultramicroscopy*, vol. 157, pp. 35–47, Oct. 2015.

[40] G. Wang, M. W. Vannier, and P.-C. Cheng, "Iterative x-ray cone-beam tomography for metal artifact reduction and local region reconstruction," *Microscopy and Microanalysis*, vol. 5, no. 1, p. 58–65, 1999.

[41] Y. Zhang, W.-H. Zhang, H. Chen, M.-L. Yang, T.-Y. Li, and J.-L. Zhou, "Few-view image reconstruction combining total variation and a high-order norm," *International Journal of Imaging Systems and Technology*, vol. 23, no. 3, pp. 249–255, Aug. 2013.

[42] B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, and R. M. S., "Image reconstruction by domain-transform manifold learning," *Nature*, vol. 555, no. 7697, pp. 487–492, Mar. 2018.

[43] S. C. Zhu, Y. Wu, and D. Mumford, "Filters, random fields and maximum entropy (frame): Towards a unified theory for texture modeling," *International Journal of Computer Vision*, vol. 27, no. 2, pp. 107–126, 1998.

SliTraNet: Automatic Detection of Slide Transitions in Lecture Videos using Convolutional Neural Networks

Aline Sindel[1], Abner Hernandez[1], Seung Hee Yang[2], Vincent Christlein[1] and Andreas Maier[1]

***Abstract*— With the increasing number of online learning material in the web, search for specific content in lecture videos can be time consuming. Therefore, automatic slide extraction from the lecture videos can be helpful to give a brief overview of the main content and to support the students in their studies. For this task, we propose a deep learning method to detect slide transitions in lectures videos. We first process each frame of the video by a heuristic-based approach using a 2-D convolutional neural network to predict transition candidates. Then, we increase the complexity by employing two 3-D convolutional neural networks to refine the transition candidates. Evaluation results demonstrate the effectiveness of our method in finding slide transitions.**

I. INTRODUCTION

Nowadays, there is a huge number of online learning material available to students and researchers. Lecture videos uploaded by the universities to video sharing platforms such as YouTube or to in-build video platforms are accessible from anywhere and at any time. The high amount of video material makes it tedious for the user to search for specific content by browsing through the individual videos. Hence, video summarization can help to quickly grasp the overview of the lecture video. This can be done by the automatic detection of slide transitions to extract the slide and time stamp at each slide change. Automatic detection of slide transitions can also support the lecturer in creating lecture notes. In combination with the audio transcript of the lecture video, the extracted slides can be automatically inserted into the audio text based on their time stamp. For instance, the free video-to-blog post conversion software AutoBlog [21] automatically extracts the transcript of a lecture video to generate a blog post [9]. So far, the slides are manually inserted into the blog text. However, using our slide transition detection method, the software could be extended.

The variety in the types of lecture videos makes the task challenging. For example, the lecture slides can be full screen with the lecturer screen inserted as a small window on top, or the lecture slides can be depicted next to the view of the lecturer. Further, memes (e.g. animations and short videos) to illustrate the lecture content can be inserted into the lecture video. Memes and the actual slides can have very similar frames from the style and color distribution. Thus, lecture videos that not only contain the slides and the speaker's view, but also these meme videos make the task even more difficult.

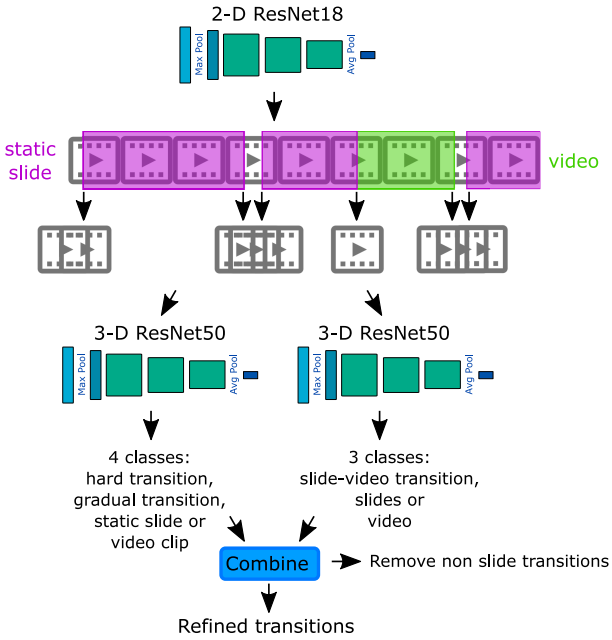


Fig. 1: Overview of our SliTraNet for slide transition detection: First, we predict initial slide-slide or slide-video transition candidates by comparing each frame (cropped to the slide content) to its respective anchor frame using a 2-D ResNet. At the transition candidate positions, we extract overlapping video clips with a length of eight frames from the cropped video and the raw video. Two 3-D ResNets have been trained to extract spatio-temporal features to classify the cropped video clips into hard or gradual transitions, static slides or video sequences and the raw video clips into slide-video transitions, slide sequences or video sequences. Lastly, we combine the class predictions of both 3-D ResNets to exclude transitions mutually classified as video sequence.

In this paper, we propose a deep learning method for the detection of slide transitions in lecture videos, which we train and test on a dataset that contains video sequences of lectures with slides, speaker views, and memes. To detect the slide transitions, we present a multi-step approach. First, we predict initial transition candidates by inserting a 2-D convolutional neural network (CNN) into a heuristic-based approach. Then, we extract spatio-temporal features at the candidate positions using two 3-D CNNs to exclude transitions that were classified as video sequences.

This work was supported by the project "MEOW" funded by "DAAD IP Digital"

[1]Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), 91058 Erlangen, Germany {aline.sindel, abner.hernandez, vincent.christlein, andreas.maier}@fau.de

[2]Department of Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), 91058 Erlangen, Germany seung.hee.yang@fau.de

II. RELATED WORK

This section summarizes related works in the field of slide transition detection, scene boundary detection, and video thumbnail selection.

A. Slide Transition Detection

Traditional approaches to slide detection focus on low-level features to measure the similarity across adjacent frames. For example, the maximum peak of the color histogram and difference in entropy for horizontal lines were used to detect slide changes in [14]. Often the use of histograms for slide detection is supplemented with other algorithms to detect features such as faces, or text [20], [29]. Similarly in [2], histograms are utilized for shot boundary detection as part of a larger scheme involving shot classification, slide region detection, and slide transition detection.

The variance in image scaling and rotation can be handled by the Scale Invariant feature transform (SIFT) algorithm. This approach detects slide transitions when the SIFT similarity is under a defined threshold. Features extracted using the SIFT algorithm have shown good slide detection accuracy rates in [10], [22] and with slide alignment [28]. SIFT features can also be used with sparse time-varying graphs [17], where the graph models slide transitions. The temporal modeling of slide transitions can also be conducted using a Hidden Markov Model (HMM), where the states of the model correspond to an individual slide [5], [24], [3], [4]. The likelihood of the states are computed with a correlation measure and the most probable sequence of slides is calculated using the Viterbi algorithm.

The current study approaches the slide transition detection problem by using 3-D CNNs which can learn spatio-temporal features that are useful for detecting slide transitions. However, the training time and memory consumption can be problematic. Therefore, Residual Networks (ResNet [8]) have been suggested by [18] for this task. They propose a novel residual block that contains an extra 1×1 3-D convolutional layer to the shortcut connection layer. They show better results for ResNet compared to the traditional slide transition approaches on their to 6 frames per second temporally downsampled dataset. In [6], a Dual Path Network (DPN) [1] that combines both ResNeXt and DenseNet is proposed. Further, they introduce a Convolutional Block Attention module to their network that sequentially infers a 1-D channel attention map, followed by a 2-D spatial attention map, and lastly a 1-D time attention map. Further improvements in the F_1-score were obtained compared to traditional approaches or with ResNets alone.

B. Scene Boundary Detection

A related field of work is scene boundary detection or shot boundary detection (SBD) [11]. Traditionally, SBD relied on the same low-level features such as histograms. However, the issue of detecting changes is complex and requires attention to the variability of transitions. Detecting gradual transitions is a particularly difficult problem and recent studies on SBD now take into consideration the presence of sharp cut transitions and gradual transitions. For example, a 3-D CNN-based model from [7] was combined with an SVM classifier to label frames as being either normal, a gradual transition, or a sharp transition. In [15], both types of transitions are detected by separate 3-D CNNs. A similar approach using deep CNNs was taken by [27] where SBD was implemented via a three stage process; candidate detection, cut transition detection, and gradual transition detection. TransNet [26] and TransNet2 [25] use Dilated DCNNs to detect sharp and gradual transitions.

C. Video Thumbnail Selection

Another related area is video thumbnail selection, which summarizes the video content by selecting a representative frame as the thumbnail. To extract the representative frames, learning-based approaches have been proposed that take the user's perspective selection of representative frames into account [12], [19]. Based on visual features, the videos are classified according to image quality, visual details, user attention, and display duration [12], or different types of camera motion [19]. Approaches also combine the visual content with side semantic information such as the title or transcript for query-dependent thumbnail selection [16] or to visually enrich the thumbnail with keywords [30].

III. METHODOLOGY

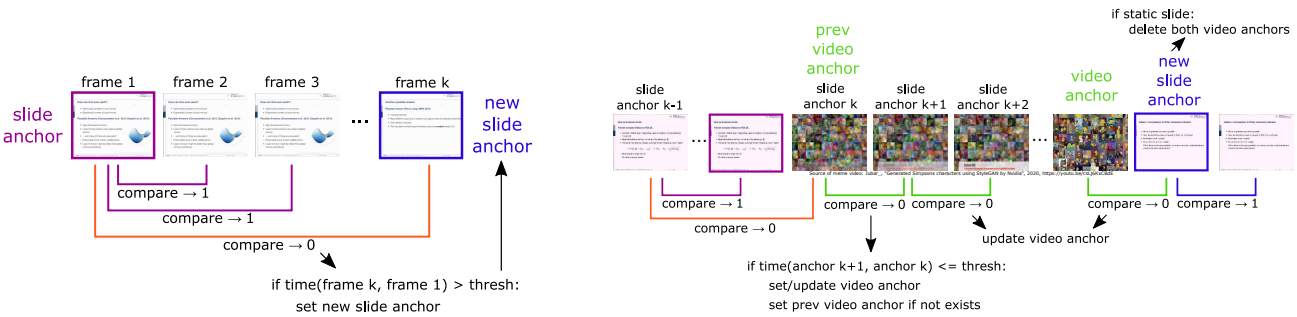
In this section, our method for slide transition detection is presented. We describe the network architectures and introduce training and inference of the different parts of the pipeline.

A. Overview of SliTraNet

SliTraNet is composed of three convolutional neural networks, which are all separately trained for the three different tasks and combined for inference, see Fig. 1. We process the complete data once by applying a 2-D ResNet18 [8] to pairs of each frame with its anchor frame resulting in initial slide-slide or slide-video transition candidates. For the refinement step, we increase the complexity of the networks by using two 3-D ResNet50s and apply these to video clips extracted from the transition candidate positions. A short video clip of eight frames can contain a sequence with one hard transition, a gradual transition, a static sequence of the same slide or a sequence of video frames, such as a short animation, a speaker view, or a meme. We train one 3-D ResNet for these four classes and another 3-D ResNet to distinguish slide-video transitions, slide sequences, and video sequences. Based on the class predictions, we exclude transition candidates that were classified as video sequence by both networks.

B. Initial Transition Candidates Estimation

We train a 2-D ResNet18 for the discrimination task whether two images are from the same slide (class 1) or not (class 0). For this task, we concatenate both images along the color channel dimension to obtain 6 channels for RGB input or 2 channels for grayscale input and modify the



(a) Detection of a static slide sequence

(b) Detection of a video frame sequence

Fig. 2: Comparison to anchor frame using a neural network to detect static slide sequences and video sequences.

input channels of the ResNet18 [8] architecture accordingly. For training, we generate the same number of positive and negative pairs. For the negative pairs, we first select frames from the neighboring slides for each slide and then fill the rest with randomly chosen frames that have a different slide id. For the optimization, we employ the binary cross-entropy loss.

To predict the transition candidates, we plug the neural network into a heuristic-based approach, as illustrated in Fig. 2a and 2b. We compare each frame to an anchor frame by the neural network to search for static slides (Fig. 2a). As long as both frames are classified as the same, we keep the anchor and as soon as the two frames are classified as different, we set the anchor to the current frame. A static slide is detected if the time, measured in number of frames, is higher than a threshold. This general idea is borrowed from Perelman [23], which uses the absolute difference of the blurred grayscale versions of the frames. Since the lecture videos also contain video sequences without slides, we extended the approach further by adding two video anchors, see Fig. 2b. If a frame difference is detected by the neural network and the time from the current frame k to the anchor $k-1$ is smaller or equal to the threshold, the video anchor and previous video anchor are set to the current frame k. As long as the frames are not classified as the same, the video anchor is updated. After the next static slide sequence is detected, a video sequence is recorded from the previous video anchor to the video anchor and both anchors are deleted. The slide-slide and slide-video transition candidates are determined from the detected static slide and video sequences.

C. Transition Candidates Refinement

Since the video sequences can also contain static frames that might be classified as static slides using the deep-heuristic-based approach, a refinement step is necessary to reduce the number of false positives. To better exploit the spatio-temporal character of the video, we train a 3-D ResNet50 using cross-entropy loss for the multi-task classification problem that assigns a short video clip of eight frames to one of the classes: hard transition, gradual transition, static slide and video. The network architecture is depicted in Fig. 3, which is slightly adapted from the 3-D ResNet backbone in [13]. For the initial layers and 3-D max pooling, we modified the strides of the temporal dimension to be 1 to only reduce the spatial dimensions.

The slides of lecture videos are not necessarily filling the full screen, but can be placed on top of some background. In our particular lecture video dataset, the memes, animations and speaker video sequences are full screen in contrast to the slides, see Fig. 4. Using this knowledge, we use the raw video input to train our second 3-D ResNet50 to classify the short clip into slide-video transitions, slide sequences or video sequences.

For training both networks, we extract video clips at striking positions such as placing the middle of the clip (plus minus one frame) at the position of the hard transition, the begin, middle and end of the gradual transition and in the middle of a static slide sequence or at some equally spaced positions within the video sequence. For the second task, the slide-video transitions occur only rarely in the dataset in comparison to slide sequences or video sequences. Hence, we use the weighted cross-entropy loss to account for the frequency of the classes.

During inference, we use the predictions of the deep-heuristic-based approach to extract the video clips to feed them to the 3-D ResNets and based on the output of both networks, we filter out slide transition candidates that were classified to be video sequences by both networks.

IV. EXPERIMENTS AND RESULTS

In this section, we describe our dataset and measure the performance of our method.

A. Lecture Video Dataset

The dataset comprises a subset of lecture videos from two courses in the field of deep learning and medical image processing of the Pattern Recognition Lab, FAU Erlangen-Nuremberg. The videos are recorded in Full HD with 25 frames per second and range between a duration of 6 to 33 min. The slides of one course are in the format $4:3$ and of the other in $16:9$. The dataset is split into 12 videos for training, 4 for validation and 14 for testing. To feed the data to the network, the frames are cropped to the content of the slides except for the video-slide differentiation task

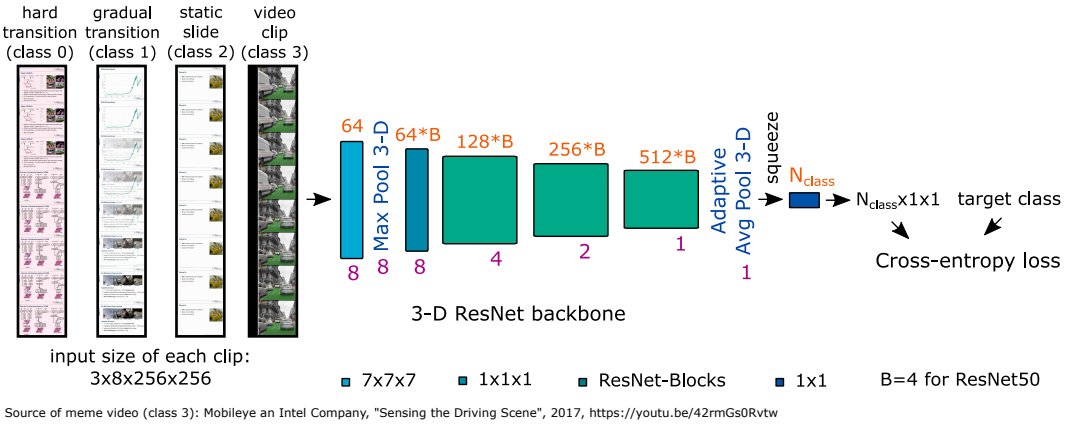


Fig. 3: Training of 3-D ResNet50 for the multi-class classification task: detection of hard transitions, gradual transitions, static slides, and videos. For each input clip one class is predicted. The numbers in orange indicate the number of output feature maps of each convolutional layer or block and the numbers in purple denote the output dimension of the temporal dimension.

Fig. 4: Frames of the lecture video dataset. Top row: raw video frames, bottom row: cropped frames.

(see Fig. 4) and for all tasks are scaled to a maximum length of 256 and filled up with zero padding to a patch size of 256×256. The ground truth slide transitions were obtained semi-automatically. Based on the difference of the frames, static slides were roughly detected and were manually corrected at frame level and split into hard and gradual transitions.

B. Implementation Details

We trained all networks from scratch for 100 epochs with early stopping using the following training parameters: learning rate $\eta = 2 \cdot 10^{-4}$, linear decay to 0 starting at epoch 50 for 2-D ResNet18 and 60 for 3-D ResNet50, Adam solver, momentum $(0.9, 0.999)$, batch size of 64 for 2-D ResNet18 and of 32 for 3-D ResNet50 for training and validation and online data augmentation for the training data split (color jittering, horizontal flipping, color inversion, Gaussian blurring with kernel size in range 1 to 21, reversed ordering of the clips and one frame offsets at clip extraction). For inference, the threshold for static slides is set to 8 frames.

C. Qualitative and Quantitative Evaluation

We evaluate our method using precision, recall, and F_1 score of slide transitions for our test dataset. Since the gradual transitions are annotated and predicted by our method as frame intervals, we compare the closest euclidean distances of the start and end points of the predicted and labeled transitions to a threshold of 20. This comparison is performed bi-directionally and the mutually valid counts determine the number of true positive transitions.

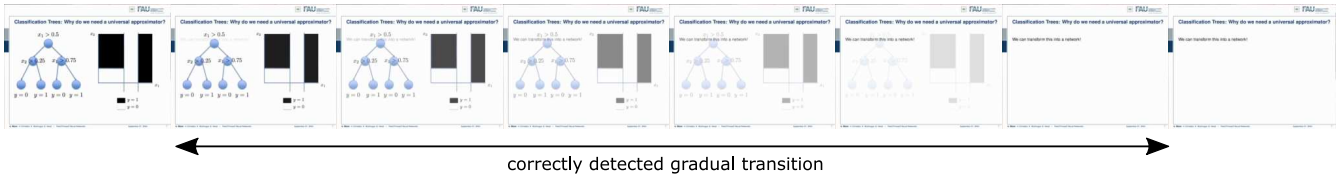
The quantitative evaluation results are summarized in Tab. I. In the top rows, we compare the first step of our approach using the 2-D ResNet18 (trained and tested in RGB and grayscale) to the traditional approach inspired by [23] of using the frame difference with Gaussian blur (kernel size $k_s = (21, 21)$) in RGB and grayscale. From these methods, the grayscale 2-D ResNet achieves the highest F_1 score, which is slightly above 50 %. The 2-D methods have a high recall but a low precision due to their high number of false positives. The frame to anchor comparison detects many false positive transitions for video frames, where short static sequences alternate with motions.

Hence, the second part of our pipeline is necessary to reduce these false positives, whose results are shown in the bottom rows of Tab. I. Using the combination of the first step and the 3-D ResNets a performance gain in the F_1 score of up to 35 % is achieved, i.e., our SliTraNet reaches an F_1 score of almost 90 %, which is closely followed by the combination of difference + 3-D ResNets. This second step maintains the high recognition rate while decreasing the number of false positives, resulting in higher precision, which is partly due to the spatio-temporal convolutions in the 3-D ResNets that recognize the different transition types better than the 2-D approach.

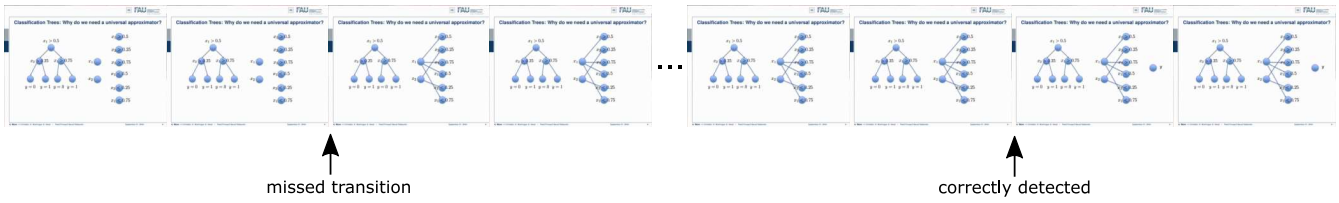
Additionally, we evaluate how the order of the networks influences the result by reversing the order. First, we apply the 3-D ResNet to classify overlapping video clips of length 8 into slide-video, slides and videos. We use the slide-video and slide candidates to apply the next 3-D ResNet to classify the remaining clips into the transition types, static slides and videos. We iterate through the potential transition regions and apply the 2-D ResNet pairwise to localize slide changes. This

TABLE I: Evaluation of precision, recall, and F_1 score of slide transition detection for the test set with 14 videos. In the top rows is the comparison of the first step of the approach: 2-D ResNet18 versus difference with Gaussian blur in both color and grayscale. In the bottom rows the combination of the above methods with the 3-D ResNet50 (in color) and the application of the three networks in reverse order (first 3-D then 2-D) is shown.

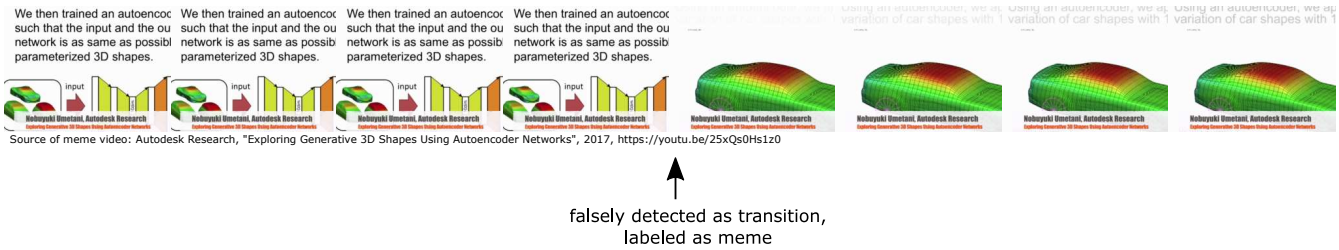
	Number of transitions	TP	FP	FN	**Precision**	**Recall**	**F_1 score**
Ground Truth	380	380	0	0	100.00	100.00	100.00
Diff-RGB-blur	992	365	627	15	36.79	**96.05**	53.21
Diff-gray-blur	1011	365	646	15	36.10	**96.05**	52.48
2-D ResNet18-RGB	1188	358	830	22	30.13	94.21	45.66
2-D ResNet18-gray	911	355	556	25	**38.97**	93.42	**55.00**
ResNets-Reverse-RGB-gray	366	303	63	77	82.79	79.74	81.23
Diff-RGB-blur + 3-D ResNet50-RGB	435	364	71	16	83.68	**95.79**	89.33
Diff-gray-blur + 3-D ResNet50-RGB	442	364	78	16	82.35	**95.79**	88.56
SliTraNet-RGB-RGB	453	357	96	23	78.81	93.95	85.71
SliTraNet-gray-RGB	408	354	54	26	**86.76**	93.16	**89.85**



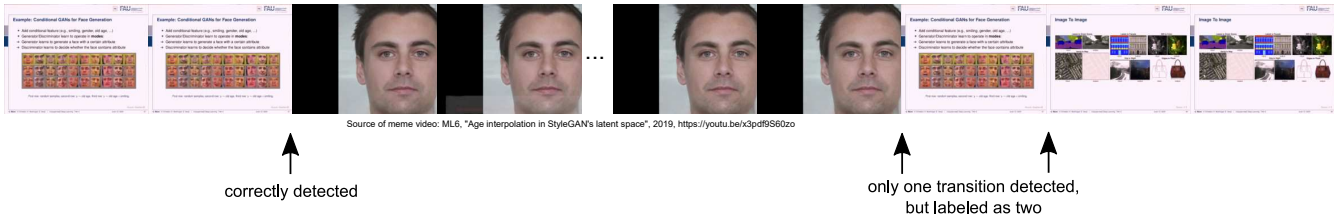
(a) A true positive gradual transition



(b) One false negative transition in an animated slide



(c) A false positive transition in case of memes



(d) One false negative transition due to fast slide change after meme insertion

Fig. 5: Qualitative results for slide transition detection using SliTraNet: Correct and failure cases.

approach with an F_1 score of around 81 % misses more slide transitions than the competing methods and due to the high complexity in the first two steps consumes a long execution time. In contrast, SliTraNet takes less than 90 min to process the 190 min test data.

Overall our SliTraNet demonstrates high effectiveness in the task of slide transition detection in lecture videos, which is also confirmed by the qualitative evaluation. In Fig. 5 some difficult cases are depicted to highlight the advantage of our method and also define some limitations. One difficulty is

represented by animated slides, where little content changes in a short time. Fig. 5a shows an example of a correctly detected gradual transition, where the start and end point are marked by the blue arrow. From the hard transitions in Fig. 5b only the right one is detected by SliTraNet. A plausible reason for the failure of the network for the first transition is the small difference of the two frames as only thin lines appear that connect the nodes, while for the detected transition the slide change is larger due to the added node. Another difficulty that arises are the memes that are inserted into the lecture videos. The meme in Fig. 5c has a similar color distribution as the lecture slides and thus the transition within the meme is falsely detected as a slide transition. In Fig. 5d an example is shown, where the meme was inserted to the end of a static slide. The slide-video transition is correctly detected, but from the two fast slide changes, only one is detected. In the first step of the approach, we defined that a static slide has to be at least eight frames long, hence slides of one frame length cannot be detected by our method, but for the most applications these limitations are acceptable.

V. CONCLUSIONS

We presented a deep learning method to detect slide changes in lecture videos such as hard and gradual transitions. The quantitative evaluation showed a high performance of our method for this task. Future work could comprise extending the approach for a larger dataset and integrating it for online teaching, for instance to automatically insert slides for creating lecture notes in the AutoBlog framework.

REFERENCES

[1] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng, "Dual Path Networks," *Advances in Neural Information Processing Systems (NIPS)*, vol. 30, pp. 4467–4475, 2017.

[2] P. Eruvaram, K. Ramani, and C. S. Bindu, "An Experimental Comparative Study on Slide Change Detection in Lecture Videos," *International Journal of Information Technology (IJIT)*, vol. 12, no. 2, pp. 429–436, 2020.

[3] Q. Fan, A. Amir, K. Barnard, R. Swaminathan, and A. Efrat, "Temporal Modeling of Slide Change in Presentation Videos," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 989–992, 2007.

[4] Q. Fan, K. Barnard, A. Amir, and A. Efrat, "Robust Spatiotemporal Matching of Electronic Slides to Presentation Videos," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2315–2328, 2011.

[5] G. Gigonzac, F. Pitie, and A. Kokaram, "Electronic Slide Matching and Enhancement of a Lecture Video," *4th European Conference on Visual Media Production (CVMP)*, pp. 1–7, 2007.

[6] M. Guan, K. Li, R. Ma, and P. An, "Convolutional-Block-Attention Dual Path Networks for Slide Transition Detection in Lecture Videos," *International Forum on Digital TV and Wireless Multimedia Communications (IFTC)*, pp. 103–114, 2019.

[7] A. Hassanien, M. Elgharib, A. Selim, S.-H. Bae, M. Hefeeda, and W. Matusik, "Large-Scale, Fast and Accurate Shot Boundary Detection Through Spatio-Temporal Convolutional Neural Networks," *arXiv preprint arXiv:1705.03281*, 2017.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[9] A. Hernandez and S. H. Yang, "Multimodal Corpus Analysis of Autoblog 2020: Lecture Videos in Machine Learning," *Proceedings of the International Conference on Speech and Computer (SPECOM)*, pp. 262–270, 2021.

[10] H. J. Jeong, T.-E. Kim, H. G. Kim, and M. H. Kim, "Automatic Detection of Slide Transitions in Lecture Videos," *Multimedia Tools and Applications*, vol. 74, no. 18, pp. 7537–7554, 2015.

[11] H. Jiang, A. S. Helal, A. K. Elmagarmid, and A. Joshi, "Scene Change Detection Techniques for Video Database Systems," *Multimedia systems*, vol. 6, no. 3, pp. 186–195, 1998.

[12] H.-W. Kang and X.-S. Hua, "To Learn Representativeness of Video Frames," *Proceedings of the Annual ACM International Conference on Multimedia*, p. 423–426, 2005.

[13] O. Köpüklü, X. Wei, and G. Rigoll, "You Only Watch Once: A Unified CNN Architecture for Real-Time Spatiotemporal Action Localization," *arXiv preprint arXiv:1911.06644*, 2019.

[14] W. H. Leung, T. Chen, F. Hendriks, X. Wang, and Z.-Y. Shae, "eMeeting: A Multimedia Application for Interactive Meeting and Seminar," *IEEE Global Telecommunications Conference (GLOBECOM)*, pp. 2994–2998, 2002.

[15] R. Liang, Q. Zhu, H. Wei, and S. Liao, "A Video Shot Boundary Detection Approach Based on CNN Feature," *IEEE International Symposium on Multimedia (ISM)*, pp. 489–494, 2017.

[16] W. Liu, T. Mei, Y. Zhang, C. Che, and J. Luo, "Multi-Task Deep Visual-Semantic Embedding for Video Thumbnail Selection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3707–3715, June 2015.

[17] Z. Liu, K. Li, L. Shen, and P. An, "Sparse Time-Varying Graphs for Slide Transition Detection in Lecture Videos," *International Conference on Image and Graphics (ICIG)*, pp. 567–576, 2017.

[18] Z. Liu, K. Li, L. Shen, R. Ma, and P. An, "Spatio-Temporal Residual Networks for Slide Transition Detection in Lecture Videos," *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 13, no. 8, pp. 4026–4040, 2019.

[19] J. Luo, C. Papin, and K. Costello, "Towards Extracting Semantically Meaningful Key Frames From Personal Video Clips: From Humans to Computers," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 19, no. 2, pp. 289–301, 2009.

[20] D. Ma and G. Agam, "Lecture Video Segmentation and Indexing," *Proceedings of the Society of Photo-Optical Instrumentation Engineers (SPIRE), Document Recognition and Retrieval XIX*, pp. 238–245, 2012.

[21] A. Maier, "AutoBlog," https://autoblog.tf.fau.de/, 2020.

[22] A. Mavlankar, P. Agrawal, D. Pang, S. Halawa, N.-M. Cheung, and B. Girod, "An Interactive Region-Of-Interest Video Streaming System for Online Lecture Viewing," *18th International Packet Video Workshop (PV)*, pp. 64–71, 2010.

[23] D. Perelman, "Slide-detector," https://git.aweirdimagination.net/perelman/slide-detector, 2020.

[24] G. Schroth, N.-M. Cheung, E. Steinbach, and B. Girod, "Synchronization of Presentation Slides and Lecture Videos Using Bit Rate Sequences," *18th IEEE International Conference on Image Processing (ICIP)*, pp. 925–928, 2011.

[25] T. Souček and J. Lokoč, "TransNet V2: An Effective Deep Network Architecture for Fast Shot Transition Detection," *arXiv preprint arXiv:2008.04838*, 2020.

[26] T. Souček, J. Moravec, and J. Lokoč, "TransNet: A Deep Network for Fast Detection of Common Shot Transitions," *arXiv preprint arXiv:1906.03363*, 2019.

[27] T. Wang, N. Feng, J. Yu, Y. He, Y. Hu, and Y.-P. P. Chen, "Shot Boundary Detection Through Multi-stage Deep Convolution Neural Network," *International Conference on Multimedia Modeling (MMM)*, pp. 456–468, 2021.

[28] X. Wang and M. Kankanhalli, "Robust Alignment of Presentation Videos with Slides," *Pacific-Rim Conference on Multimedia (PCM)*, pp. 311–322, 2009.

[29] B. Zhao, S. Lin, X. Qi, R. Wang, and X. Luo, "A Novel Approach to Automatic Detection of Presentation Slides in Educational Videos," *Neural Computing and Applications*, vol. 29, no. 5, pp. 1369–1382, 2018.

[30] B. Zhao, S. Lin, X. Qi, Z. Zhang, X. Luo, and R. Wang, "Automatic Generation of Visual-Textual Web Video Thumbnail," *Special Interest Group on Computer Graphics and Interactive Techniques (SIGGRAPH) Asia Posters*, pp. 1–2, 2017.

Data Synthesis for Large-scale Supermarket Product Recognition

Julian Strohmayer[1] and Martin Kampel[1]

***Abstract*— Training data acquisition for deep learning-based visual product recognition systems on a large scale is laborious and often infeasible due to the vast product assortments containing tens of thousands of products and the densely packed scenes. In this work, we propose a potential solution to this problem in the form of an automatic data synthesis pipeline that can generate training data for product detectors and classifiers on a large scale. To demonstrate that our synthesis pipeline can produce realistic data, we train a product detector using only synthetic data and measure its generalization to real data. A detection accuracy of 0.832 mAP@0.50 is achieved on real data, showing that the model can learn from our synthetic data.**

I. INTRODUCTION

Visual product recognition is a contemporary computer vision problem where the aim is the detection and classification of individual products in images of supermarket environments [19]. Potential applications of visual product recognition include automatic checkout systems [18], real-time inventory management [3], planogram compliance [13] or assistive technologies for the visually impaired [8]. As with generic visual object recognition, deep learning models have proven effective in the special case of visual product recognition. The amount of labeled data required for training is, however, not easily acquired. The vast and constantly changing product assortments, covering tens of thousands of products, make manual acquisition and labeling infeasible. A promising solution to this data acquisition problem are synthesis methods [14] that can automatically generate practically unlimited amounts of training data with pixel-accurate labels. While promising, the synthesis of realistic data is challenging and requires a thorough understanding of the target domain since any domain gap can limit model generalization. In this paper, we present a scalable data synthesis pipeline for the problem of supermarket product recognition capable of generating realistic training data for product detectors and classifiers.

II. RELATED WORK

In [19], Wei et al. conduct a comprehensive survey on the current state of visual product recognition, which discusses both challenges and techniques. A recent work by Qiao et al. [15] proposes a synthesis approach similar to ours to investigate the problem of object proposal generation in supermarket images. A virtual supermarket with 1438 3D product models is built in the Unreal Engine, allowing the generation of randomized supermarket shelves. A synthetic dataset, generated with the proposed method, is used in combination with the MS COCO dataset [12] to train a product detector, showing that the use of synthetic data improves detection accuracy. As demonstrated, the use of 3D models allows the generation of very realistic images. However, this approach does not scale well to tens of thousands of products, which is why we choose to approximate complex product geometries with billboards. Furthermore, in [3], Follmann et al. synthesize training data for the problem of supermarket product instance segmentation by randomly recombining segmented products. The authors demonstrate that the introduction of synthetic data greatly improves the detection and segmentation accuracy of Mask R-CNN [6], FCIS [9], Faster R-CNN [16] and RetinaNet [10] models. A recent development is the use of generative models for product image synthesis. In [18], Wei et al. generate a synthetic checkout dataset using CycleGAN [20]. Another work in this direction by Tonioni et al. [17] employs a GAN [5] to synthesize realistic product images for the training of an embedder model.

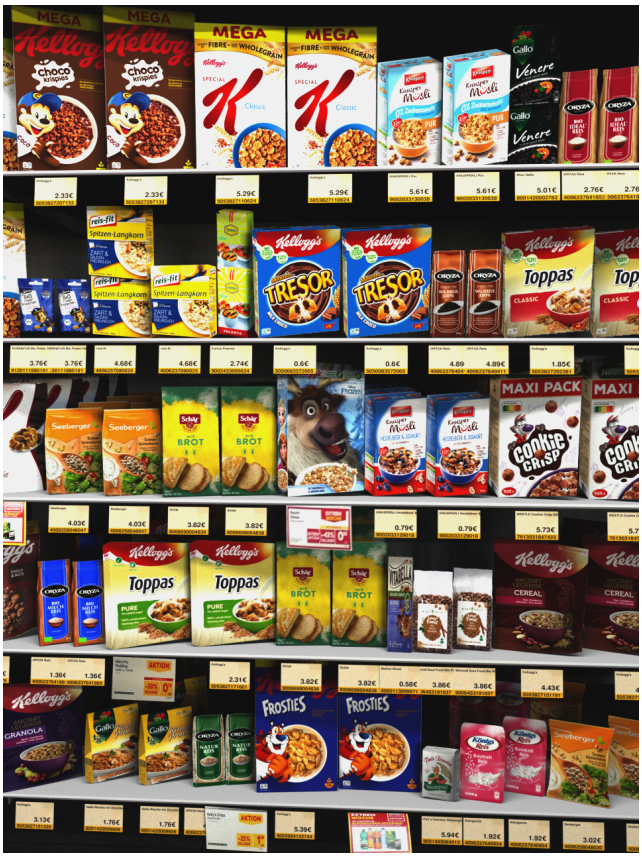


Fig. 1: Synthetic shelf image with cereal products, generated with the proposed data synthesis pipeline.

*This work is funded by the Wirtschaftsagentur Wien under grant 3540290.

[1]TU Wien, Computer Vision Lab, 1040 Vienna, Austria. {julian.strohmayer,martin.kampel}@tuwien.ac.at

III. DATA SYNTHESIS

Our data synthesis pipeline is based on the free and open-source 3D creation suite Blender[1], which we control with a Python script to generate randomized scenes of supermarket shelves. The input data for our synthesis pipeline is sourced from a commercial product database of GS1 Austria[2], covering the Austrian market. Relevant fields are Global Trade Item Number (GTIN), product image, and product dimensions. This information is passed to the synthesis pipeline in the form of a product list which automatically generates a suitable training dataset. For a single data sample, the synthesis process is as follows. First, an empty shelf is created according to predefined parameters such as shelf height, shelf depth, and the number of levels. Then each level of the shelf is then randomly populated with products sampled from the same product family to prevent unrealistic product combinations (e.g., cereals and dairy products). For this, the Global Product Classification (GPC) system is used to classify the products at the GPC family level. To ensure that each product is present at least once in the generated dataset, a target product is selected for each image, which is iteratively extracted from the product list. The target product is placed, clearly visible, in the shelf center. The camera position is then randomly chosen within a hemisphere of radius $r \in [0.75m, 2m]$ in front of the target product. As we only have a single frontal image per product, complex 3D product geometry is approximated as billboards with a locked horizontal rotation axis. We map alpha textures onto the billboards and align them relative to the camera position, creating the impression of 3D geometry. After texturing, the scene is rendered using the Cycles renderer and saved as 8bit RGB image. For the generation of the bounding box labels, each product is assigned a unique Blender object ID, and an object ID pass of the scene is performed. The resulting binary masks for each product are used to calculate the bounding boxes. To improve label quality, bounding boxes smaller than 0.1% of the image size are eliminated. GTIN, class labels and normalized bounding box coordinates (x_c, y_c, w, h) of all products in the scene are combined in a separate label file. The rendering of a 960×1280 image with 512 antialiasing samples and the generation of the corresponding labels takes 120 seconds on an Nvidia RTX 2070 GPU.

IV. EVALUATION

To evaluate whether the proposed synthesis pipeline can generate realistic training data for the problem of product recognition, a deep learning model is trained exclusively on synthetic data (no pretraining or finetuning on real data), and its generalization to real data is assessed. We quantify the domain gap between synthetic and real data by measuring detection accuracy simultaneously on a synthetic and a real validation dataset during training.

[1]Blender, https://www.blender.org/, accessed: 24.09.2021
[2]GS1 Austria, https://www.gs1.at/, accessed: 24.09.2021

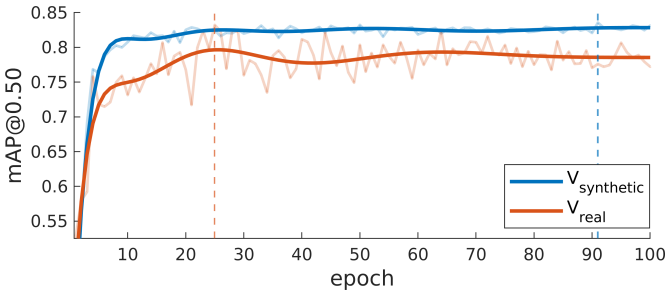


Fig. 2: Measured detection accuracy on $V_{synthetic}$ and V_{real} over 100 training epochs.

A. Data

For model training, a synthetic dataset of 1000 images, composed of products of the GPC families 50100000, 50220000, 50200000, 50120000, and 50190000, is synthesized with the proposed method. An example image is given in Fig. 1. The 1000 images are further split randomly in a 9:1 ratio into a training and validation dataset $V_{synthetic}$. The real validation dataset V_{real}, used to assess generalization to real data, consists of 100 real supermarket shelf images, which were captured in three different Viennese supermarkets and annotated by hand.

B. Model Training

As network architecture for this evaluation, we use a RetinaNet [11] with ResNet50 [7] backbone, implemented in the PyTorch torchvision.models[3] package. The model is trained as a class agnostic product detector that distinguishes between two classes (product and background), as the underlying class number would exceed the capabilities of current monolithic classifiers [4][19]. To erase any prior knowledge derived from real data (MS COCO or ImageNet [1]) that could interfere with our measurements, the model is trained from scratch (*pretrained=False, pretrained_backbone=False, trainable_backbone_layers=5*). Predicted bounding boxes with excessive overlap are eliminated by choosing a non-maximum suppression threshold of 0.15 Intersection over Union (IoU). The model is trained for 100 epochs using the Adam optimizer with exponential learning rate decay from 0.0001 to 0.00001 and a batch size of 1. Detection accuracy is measured as PASCAL VOC [2] mean Average Precision (mAP) at 0.5 IoU, which we denote as mAP@0.50 hereafter.

C. Results

The training progress of our product detector model is visualized in Figure 2, showing the mAP@0.50 on $V_{synthetic}$ and V_{real} over 100 training epochs. We achieve a maximum mAP@0.50 of 0.836 and 0.832 on $V_{synthetic}$ and V_{real}, respectively. While a small domain gap can be observed over the training period, the strong correlation between the datasets shows that the model can generalize from synthetic to real data. At the same time, this shows that our synthesis pipeline is capable of generating realistic training data for the problem of supermarket product recognition.

[3]torchvision.models, https://pytorch.org/vision/stable/models.html, accessed: 24.09.2021

REFERENCES

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[2] M. Everingham, L. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338, 2009.

[3] P. Follmann, T. Böttger, P. Härtinger, R. König, and M. Ulrich, "Mvtec d2s: Densely segmented supermarket dataset," *ArXiv*, vol. abs/1804.08292, 2018.

[4] E. Goldman and J. Goldberger, "Crf with deep class embedding for large scale classification," *Computer Vision and Image Understanding*, vol. 191, 2019.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[6] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask r-cnn," *IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[8] M. Klasson, C. Zhang, and H. Kjellström, "A hierarchical grocery store image dataset with visual and semantic labels," *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 491–500, 2019.

[9] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, "Fully convolutional instance-aware semantic segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollr, "Focal loss for dense object detection," *IEEE International Conference on Computer Vision (ICCV)*, pp. 2999–3007, 2017.

[11] T.-Y. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2999–3007, 2017.

[12] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," *European Conference on Computer Vision*, 2014.

[13] M. Marder, S. Harary, A. Ribak, Y. Tzur, S. Alpert, and A. Tzadok, "Using image analytics to monitor retail store shelves," *IBM Journal of Research and Development*, vol. 59, no. 2/3, pp. 3:1–3:11, 2015.

[14] S. I. Nikolenko, *Synthetic Data for Deep Learning*. Springer International Publishing, 2021.

[15] S. Qiao, W. Shen, W. Qiu, C. Liu, and A. Yuille, "Scalenet: Guiding object proposal generation in supermarkets and beyond," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1809–1818, 2017.

[16] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1137–1149, 2015.

[17] A. Tonioni and L. Di Stefano, "Domain invariant hierarchical embedding for grocery products recognition," *Computer Vision and Image Understanding*, vol. 182, 2019.

[18] X.-S. Wei, Q. Cui, L. Yang, P. Wang, and L. Liu, "Rpc: A large-scale retail product checkout dataset," *ArXiv*, vol. abs/1901.07249, 2019.

[19] Y. Wei, S. N. Tran, S. Xu, B. Kang, and M. Springer, "Deep learning for retail product recognition: Challenges and techniques," *Computational Intelligence and Neuroscience*, vol. 2020, 2020.

[20] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251, 2017.

Benign Object Detection and Distractor Removal in 2D Baggage Scans

Anna Sebernegg[1] and Walter G. Kropatsch[2]

Abstract— **Baggage screening contributes to security by helping to identify threats. However, the complexity of X-ray scans and the high intra-class variability make universal appearance-based threat detection difficult. Consequently, baggage inspection still relies on human operators, and further developments to assist them in their visual search tasks are desirable. This work proposes utilizing object detection as a diagnostic aid, where distractive benign objects are automatically detected and removed from the images through inpainting. The applied distractor removal successfully reduces visual saliency in benign regions and decreases the overall clutter of the scans.**

I. INTRODUCTION

Baggage inspection is increasingly automated, especially liquid and explosive detection systems have emerged [19]. Nevertheless, automatic appearance-based threat detection is hardly available due to the challenging nature of 2D baggage scans [1], which include high levels of clutter and overlapping objects due to tightly packed luggage [21], in- and out-of-plane rotations [5], and schemes to conceal prohibited items [11]. Therefore, human operators are still required to detect threats over visual search [10]. This task demands sustained attention over extended periods [19] and is negatively affected by several factors, such as the stressful environment [14] and complexity of baggage scans [18].

One possible way to support screeners and improve the visual search task could be to enhance the baggage scans by utilizing automatic object detection as a diagnostic aid. Like computer-aided detection systems used in the medical field [9], detected regions could be processed to focus the viewer's attention on critical content that requires further investigation. One potential application proposed in this work is to reduce the number of benign items that negatively contribute to the visual clutter by detecting and inpainting them automatically.

II. RELATED WORK

Extensive research is being conducted in appearance-based object detection within baggage scans [8]. However, the focus is on threats rather than benign items. *Image processing* is another broad field utilized in scans to improve readability [16], e.g., using material filters such as the organic-only filter mentioned by Michel et al. [16]. Saliency-driven image manipulation techniques such as distractor removal [6] or attention retargeting [12] are especially explored for photography and are less common in baggage screening.

*This work was not supported by any organization

[1,2]Pattern Recognition and Image Processing group, Technische Universität Wien, Favoritenstrasse 9-11, Vienna, Austria
[1]Email: e1526184@student.tuwien.ac.at
[2]Email: krw@prip.tuwien.ac.at

III. METHODOLOGY

This paper presents and experimentally evaluates a concept for automatic detection and removal of benign items from baggage scans. The goal is to reduce distractors to diminish visual clutter and shift saliency to other image regions. Therefore, object detection and distractor removal through inpainting is applied, as visualized in Fig. 1. The inpainting method should meet the following requirements:

- Reduce the overall visual clutter of the image
- Decrease salience in the inpainted benign regions
- Maintain or even increase salience in the rest of the image, especially in regions containing threats

Before applying inpainting, the regions of interest must be identified, which is done automatically by using a Convolutional Neural Network (CNN) for object detection that provides bounding boxes of the detected threats and benign items to a subsequent semantic segmentation performed in MATLAB. The object detector is received by applying transfer learning to the pre-trained EfficientDet model (D1) provided by the TensorFlow Object Detection API [3]. The database used for training and evaluation of the model consists of 3721 X-ray scans obtained from the public mono-energy X-ray database *GDXray* [15] and baggage scans created in cooperation with the CT Research Group at Wels Campus in upper Austria. The database is divided into a training set with 2938 images, a validation set with 632 images, and a test set with 151 images, whereby no images of single objects are included for testing. The final model can detect four threats and eight benign objects. Segmentation is performed by binarizing the image using MATLAB's implementation of Otsu's method [17] and morphological operations.

The following inpainting approaches are tested to find a suitable method for distractor removal, where the last three are provided functions by MATLAB [13]: *Uniform Inpainting* (inpainting with a uniform color from the background of the image), *Inpaint Coherent* (coherence transport based inpainting as described by Bornemann and März [2]), *Inpaint Exampler* (exemplar-based inpainting as described by Criminisi et al. [4]), and *Regionfill* (inpainting by inward interpolation from the outer pixels of the region [13]).

The effects of distractor removal on human visual attention are evaluated in two ways. Firstly, quadtree complexity as proposed by Jégou and Deblonde [20] is used to obtain the enhanced baggage scan's total visual clutter and compare it to the original image to determine if the distractor removal successfully reduces clutter. This method performs quadtree decomposition, where the number of cells in the result-

ing quadtree determines the global clutter value. Secondly, local saliency changes are measured using the Itti-Koch-Niebur Saliency Model (IKN) [7] provided by the *Saliency Model Implementation Library for Experimental Research* (SMILER) [22]. A saliency map, as shown in Fig. 2, is calculated for both the original and enhanced image and compared to assess how salience is affected by the distractor removal. A secondary objective is to evaluate the effect of distractor removal on object detection by feeding the enhanced images back as input to the CNN, creating a feedback loop. CNN's, salience models, and the human primary visual cortex consider basic features, such as edges, for their computations [7]. Since distractor removal reduces these features to decrease salience in the target region, it is interesting to investigate how it affects object detection. The inpainting methods are evaluated with the test set, ignoring images without benign objects, resulting in a set of 130 images in total.

Fig. 1. From left to right: original scan, detected benign and threat objects, and distractor removal applied with *Regionfill*.

IV. EXPERIMENTAL EVALUATION

All methods except *Uniform Inpainting* reduce the total visual clutter for at least 70% of the 130 images. *Regionfill* even achieves an efficiency of 95.4%.

Uniform Inpainting is the only method that fails to decrease distractors' salience and even increase it on average. The other methods successfully reduce the salience of distractors. *Regionfill* performs best by reducing salience in 96.9% of the 130 evaluated images by an average of 31.94%. These results are consistent with the clutter measurements. The salience of threat regions is maintained or even slightly increased by the distractor removal methods.

Since the primary goal of distractor removal methods is to mask distracting elements, inpainted benign items should no longer be detectable. Therefore, a crucial question is whether the detection model correctly rejects the removed benign items. *Regionfill* performs best with 86.51% correct rejections of the 583 inpainted benign items. This result, however, also means that at least 13.49% of the removed benign items are still detectable by the model. This can be partly explained by benign objects overlapping threats, as they cannot be fully inpainted. Otherwise, there is a risk that the threat will become unrecognizable. Moreover, semantic segmentation may fail when benign and threats are close together. *Uniform Inpainting* performs worst with a correct-rejection rate of 66.27%, presumably because the shapes of the inpainted items are still very prominent, which can be seen in the rightmost image in Fig. 3. All distractor removal methods lead to an increase in true-positive detections of benign items, indicating that the model can identify additional items after distractor removal is applied. On the original images, 583 of 1015 benign items could be detected successfully. After applying the methods to the images, the model detected further 1.4%-3.2% benign items, raising the true-positive rate for benign items from 58.4% to 59.8% - 61.6%. Furthermore, false-positive detections are decreased, at most by 28.87%.

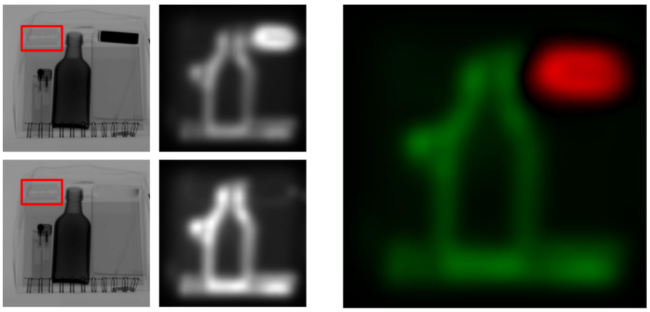


Fig. 2. From left to right: The original image with a marked threat and its enhanced version, their corresponding salience maps, and the comparison of the two maps. Red denotes a reduction in salience, and green an increase.

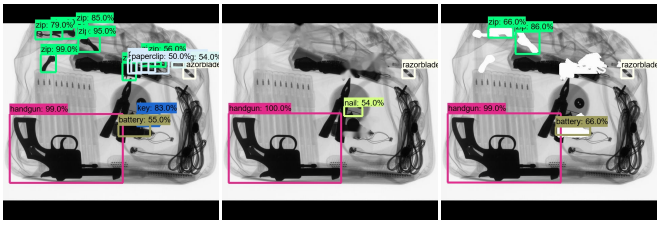


Fig. 3. Predictions on the original scan and after *Regionfill* and *Uniform Inpainting* are applied. *Regionfill* leads to a higher confidence and a new detection, while *Uniform Inpainting* fails to mask all distractors.

V. CONCLUSION

The experiments demonstrate that removing distracting items positively influences the scans' salience and successfully reduces visual clutter. The results suggest that detecting benign items in combination with distractor removal methods facilitates the visual search task, as clutter and salience are influential factors. This assumption is supported by the positive effect of the distractor removal on the object detection. Moreover, supplemental detection of benign objects provides additional information and is probably more practical than sole threat detection. While threats must not be overlooked, benign detection must only be accurate, without the need to detect all present benign objects. Furthermore, benign items are not usually deliberately concealed.

An essential disadvantage of distractor removal in 2D is that image information is artificially altered or removed without revealing new information to the viewer. Therefore, removing objects could be misleading. We hypothesize that this disadvantage is omitted as soon as more information about the bag is available, such as when working with

computed tomography that provides volumetric data of the bag. Distractor removal techniques still have to be applied with caution. For example, removing objects that are part of more complex constructions can have an undesirable effect by making the entire construction unrecognizable.

REFERENCES

[1] M. Baştan, "Multi-view object detection in dual-energy x-ray images," *Machine Vision and Applications*, vol. 26, no. 7, pp. 1045–1060, 2015.

[2] F. Bornemann and T. März, "Fast image inpainting based on coherence transport," *Journal of Mathematical Imaging and Vision*, vol. 28, no. 3, pp. 259–278, 2007.

[3] C. Chen, X. Du, L. Hou, J. Kim, P. Jin, J. Li, Y. Li, A. Rashwan, and H. Yu, "Tensorflow official model garden," 2020. [Online]. Available: https://github.com/tensorflow/models/tree/master/official

[4] A. Criminisi, P. Pérez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Transactions on image processing*, vol. 13, no. 9, pp. 1200–1212, 2004.

[5] T. Franzel, U. Schmidt, and S. Roth, "Object detection in multi-view x-ray images," in *Joint DAGM (German Association for Pattern Recognition) and OAGM Symposium*. Springer, 2012, pp. 144–154.

[6] O. Fried, E. Shechtman, D. B. Goldman, and A. Finkelstein, "Finding distractors in images," in *Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[7] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[8] D. K. Jain *et al.*, "An evaluation of deep learning based object detection strategies for threat object detection in baggage security imagery," *Pattern Recognition Letters*, vol. 120, pp. 112–119, 2019.

[9] R. T. Kneusel and M. C. Mozer, "Improving human-machine cooperative visual search with soft highlighting," *ACM Transactions on Applied Perception (TAP)*, vol. 15, no. 1, pp. 1–21, 2017.

[10] X. Liu, A. Gale, and T. Song, "Detection of terrorist threats in air passenger luggage: Expertise development," in *2007 41st Annual IEEE International Carnahan Conference on Security Technology*. IEEE, 2007, pp. 301–306.

[11] Q. Lu, *The utility of X-ray dual-energy transmission and scatter technologies for illicit material detection*. Virginia Polytechnic Institute and State University, 1999.

[12] V. A. Mateescu and I. V. Bajic, "Visual attention retargeting," *IEEE MultiMedia*, vol. 23, no. 1, pp. 82–91, 2015.

[13] MATLAB, *version 9.8.0.1380330 (R2020a)*. Natick, Massachusetts: The MathWorks Inc., 2020.

[14] J. S. McCarley, A. F. Kramer, C. D. Wickens, E. D. Vidoni, and W. R. Boot, "Visual skills in airport-security screening," *Psychological science*, vol. 15, no. 5, pp. 302–306, 2004.

[15] D. Mery, V. Riffo, U. Zscherpel, G. Mondragón, I. Lillo, I. Zuccar, H. Lobel, and M. Carrasco, "Gdxray: The database of x-ray images for nondestructive testing," *Journal of Nondestructive Evaluation*, vol. 34, no. 4, p. 42, 2015.

[16] S. Michel, S. Koller, M. Ruh, and A. Schwaninger, "The effect of image enhancement functions on x-ray detection performance," in *Proceedings of the 4th International Aviation Security Technology Symposium*, 11 2006.

[17] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[18] A. Schwaninger, S. Michel, and A. Bolfing, "Towards a model for estimating image difficulty in x-ray screening," in *Proceedings 39th Annual 2005 International Carnahan Conference on Security Technology*. IEEE, 2005, pp. 185–188.

[19] Y. Sterchi and A. Schwaninger, "A first simulation on optimizing eds for cabin baggage screening regarding throughput," in *2015 International Carnahan conference on security technology (ICCST)*. IEEE, 2015, pp. 55–60.

[20] G. Touya, B. Decherf, M. Lalanne, and M. Dumont, "Comparing image-based methods for assessing visual clutter in generalized maps," *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 2, pp. 227–233, 2015.

[21] D. Turcsany, A. Mouton, and T. P. Breckon, "Improving feature-based object recognition for x-ray baggage security screening using primed visualwords," in *2013 IEEE International Conference on Industrial Technology (ICIT)*. IEEE, 2013, pp. 1140–1145.

[22] C. Wloka, T. Kunić, I. Kotseruba, R. Fahimi, N. Frosst, N. D. Bruce, and J. K. Tsotsos, "Smiler: Saliency model implementation library for experimental research," *arXiv preprint arXiv:1812.08848*, 2018.

Explaining YOLO: Leveraging Grad-CAM to Explain Object Detections

Armin Kirchknopf[1], Djordje Slijepčević[1], Ilkay Wunderlich[2], Michael Breiter[2],
Johannes Traxler[2], and Matthias Zeppelzauer[1]

***Abstract*— We investigate the problem of explainability for visual object detectors. Specifically, we demonstrate on the example of the YOLO object detector how to integrate Grad-CAM into the model architecture and analyze the results. We show how to compute attribution-based explanations for individual detections and find that the normalization of the results has a great impact on their interpretation.**

I. INTRODUCTION

Today's complex computer vision models require mechanisms that explain their behavior. This has fueled intensive research in eXplainable Artificial Intelligence (XAI) [1]. Most work on XAI in the visual domain focuses on explaining visual classifiers, i.e., their representations learned and/or their decisions. Currently, there is a lack of XAI approaches for visual object detectors, because their special architectures impede the application of XAI methods.

In this paper, we investigate the problem of XAI for visual object detectors on the example of the YOLO detector [5]. We integrate Grad-CAM [7] into the model to generate explanations for individual object detections, i.e., bounding boxes. We compute attention maps at detection level to assess which information leads to a certain decision. For this purpose, we focus on both scores estimated by the YOLO detector, namely *objectness* and *class probability*, to obtain a more comprehensive explanation. We critically analyze the results and propose different normalization strategies to make the attention maps of different object detections within an input image or across different images comparable. We analyze results obtained for true and false detections and compare different normalization variants for result presentation.

There is a large corpus of related work both on object detection [3] and on XAI [1]. Surprisingly, the combination of both fields has hardly been investigated. Rare exceptions are the work of Tsunakawa et al. [8], who proposed an extension of a propagation-based XAI method (Layer-wise Relevance Propagation, LRP) for Single Shot MultiBox Detectors and Petsiuk et al. [4], who proposed a post-hoc model-agnostic XAI method for object detectors based on randomized input sampling. The lack of literature may be a consequence of the highly specific architectures of object detectors that impedes the integration of XAI methods. Object detectors require the explanation of localization and classification aspects and provide multiple scores that influence the likelihood of a detection. This makes the direct application of many, especially self-learned explainability approaches [2], difficult. More promising candidates are post-hoc XAI approaches. A popular example is LIME [6], which can be adapted easily to explain the final output of a detector. To explain internal scores, the direct application is, however, not possible. Additionally, the iterative probing approach of LIME makes it slow. A faster and more flexible approach is Grad-CAM, which propagates back the activation of a certain neuron (an output neuron or some intermediate neuron) to the last feature map of the underlying convolutional filter stack and uses it to weight its activations. The weighted activations in the last feature map can be directly up-scaled and overlaid with the input image to obtain an attribution-based explanation in terms of the high-level features learned by the convolutional filter stack. Note that this is more meaningful than back-propagating along the gradients completely through the network until the input pixels (guided Grad-CAM), as individual pixels lack semantic meaning.

II. METHOD

Our detection model is based on Tiny YOLO v3 [5] architecture with optimizations for inference on re-configurable hardware [9] and contains two detection heads to account for objects with different scales. The last convolutional layer of each head stores multiple scores for each potential bounding box: (i) *objectness*, which provides the likelihood for observing an object in general and (ii) a vector of *class probabilities* for all target classes. For head 1, this layer has a size of 1x1x512x30 and for head 2 1x1x256x30. Specific neurons in these layers represent the input to Grad-CAM for the generation of explanations. After these layers the YOLO architectures applies a non-maximum suppression (NMS) and a decision threshold filters out the most likely detections.

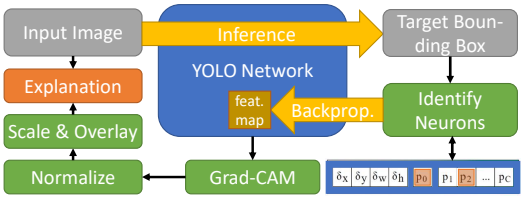


Fig. 1. Proposed explainability approach. Here p_0 represents the objectness neuron and p_2 the target class neuron in the last layer of the respective detection head.

Grad-CAM was originally proposed for conventional CNN architectures to explain decisions in terms of abstract features learned in the last convolutional layer. Considering that

*This research was funded by the Austrian Research Promotion Agency (FFG) project 876468 "SAiEX", https://bit.ly/saiex

[1]Institute of Creative Media Technologies, St. Pölten University of Applied Sciences, Austria, firstname.lastname@fhstp.ac.at.
[2]EYYES GmbH, Austria, firstname.lastname@eyyes.com.

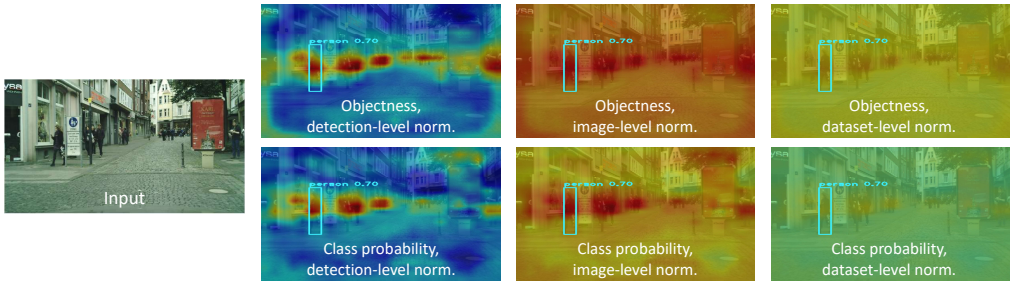


Fig. 2. Explanations for a true positive detection for objectness and class probability and three different normalization variants.

YOLO is based on a convolutional filter stack, Grad-CAM is applicable, however, not without certain modifications. For a given detection, we first identify the neurons in the last convolutional layer of the respective head corresponding to the class probability and objectness of the investigated bounding box by reversing the NMS process. These neurons represent the starting points to calculate gradients towards the neurons of the underlying convolutional layer (i.e., the top-level feature map of the convolutional stack). We follow a two-step approach to obtain explanations for both scores. The gradients are first used to weight the activation map of the underlying convolutional layer. The weighted activation map is then averaged over all channels of the layer and upscaled (i.e., interpolation) and mapped (i.e., color coding) to the input image (416px x 416px), see Figure 1. The upscaled activation pattern highlights sections in the input image that have a strong relation to the class or objectness of the investigated bounding box. Note that due to the architecture of YOLO the result of Grad-CAM are activations at the global image level, i.e., they are not limited to the observed bounding box, e.g., as shown in Figure 2.

Grad-CAM activations are by default min-max normalized to improve visibility. This leads to incomparable activations patterns between different object detections in the *same* image and across *different* images. To account for this, we propose three different normalization levels: detection-level (default), image-level (joint normalization of all explanations in an image), and dataset-level (joint normalization of all explanations across a set of images).

III. EXPERIMENTS AND RESULTS

a) Model training: The network was trained on data of front collision and rearview cameras from both public datasets including COCO, KITTI, BDD, and OpenImages as well as non-public data from the company EYYES GmbH (`www.eyyes.com`). The network was trained to detect five classes, i.e., person, cycle, car, truck, and train.

b) Experimental Setup: Our evaluation scenario originates from autonomous driving. For the evaluation, we use a subset of the Cityscapes (`www.cityscapes-dataset.com/`) dataset (which was not used for training). It consists of 3470 images showing urban street scenes from 21 cities and containing annotations of 30 classes at pixel-level. We use a subset of the above mentioned five classes. For the different normalization strategies we use min-max normalization of the Grad-CAM activations at different levels.

c) Results: Results are shown as differently normalized heatmaps overlaid on the input image for the objectness and the probability of the detected class. Figure 2 shows a correct detection of a person. The objectness shows a different activation pattern than the class probability. While class probability provides strong activations mostly on persons (including the detected one), objectness activates on all regions where the network sees potential objects. Results for detection-level normalization are most distinct, which can, however, lead to wrong conclusions, especially, when the explanation shall be compared with other detections. It actually depends on the question investigated which type of normalization is best suited. For example by normalization at dataset-level the activation strength of the detected person becomes directly comparable to all explanations in all other images and thus direct comparisons become possible which cannot be performed at detection level. This can help to develop a deeper understanding of the detector's behavior.

Figure 3 shows a falsely detected truck on the same input image. The white rectangular shaped text on the red poster seems to mislead the detector into seeing a truck. Both objectness and class probability strongly activate at detection-level raising the impression that the detector fails with high confidence. This is actually not true, which can be seen via normalization at dataset-level (not shown) where both activations are strongly attenuated, showing that the detector is actually not sure about the detection.

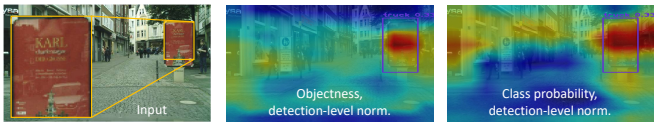


Fig. 3. Explanations for a false positive detection.

IV. CONCLUSION AND FUTURE WORK

We have investigated explainability for object detection by integrating Grad-CAM into YOLO. We can visualize its internal decision scores and thereby help to explain object detections. Our results show that normalization is essential to make different explanations comparable, e.g., across different images. Our approach is efficient: generating one explanation takes approx. half a second. In future, we aim to use these explanations to identify potential false detections at run-time.

REFERENCES

[1] Adadi and others, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, 2018.

[2] J. Chen, L. Song, M. Wainwright, and M. Jordan, "Learning to explain: An information-theoretic perspective on model interpretation," in *International Conference on Machine Learning*. PMLR, 2018, pp. 883–892.

[3] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *International journal of computer vision*, vol. 128, no. 2, pp. 261–318, 2020.

[4] V. Petsiuk, R. Jain, V. Manjunatha, V. I. Morariu, A. Mehra, V. Ordonez, and K. Saenko, "Black-box explanation of object detectors via saliency maps," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 443–11 452.

[5] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, p. 336–359, Oct 2019. [Online]. Available: http://dx.doi.org/10.1007/s11263-019-01228-7

[8] H. Tsunakawa, Y. Kameya, H. Lee, Y. Shinya, and N. Mitsumoto, "Contrastive relevance propagation for interpreting predictions by a single-shot object detector," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–9.

[9] I. Wunderlich, B. Koch, and S. Schönfeld, "An overview of arithmetic adaptations for inference of convolutional neural networks on reconfigurable hardware," in *The Sixth International Conference on Big Data, Small Data, Linked Data and Open Data*, 2019.

Case Study: Ensemble Decision-Based Annotation of Unconstrained Real Estate Images

Miroslav Despotovic[1], Zedong Zhang[1], Eric Stumpe[2] and Matthias Zeppelzauer[2]

***Abstract*— We describe a proof-of-concept for annotating real estate images using simple iterative rule-based semi-supervised learning. In this study, we have gained important insights into the content characteristics and uniqueness of individual image classes as well as essential requirements for a practical implementation.**

I. INTRODUCTION

The annotation of unlabeled images is an important task for the assignment of metadata, which can be particularly challenging within a given knowledge domain. Thus, image metadata is being increasingly used in real estate research, e.g., for valuation [9], location analysis [5], or for estimating the condition of a building [4]. In the scientific literature, there are very few contributions on the classification of unlabeled images in the domain of real estate [7]. In this short paper, we present an approach to semi-supervised labeling of images containing interior and exterior views of real estate using simple ensemble classification rule.

II. PROBLEM STATEMENT

To maximize the information potential of the data, it must be tagged with meaningful labels, which in practice can require considerable manual effort. A typical approach for annotating unlabeled data autonomously is semi-supervised learning (SSL), where an initial training set of labeled data $\mathcal{D}_\iota$ is defined by clustering and/or manual selection and the trained model is used to infer unlabeled data $\mathcal{D}_\cup$ systematically without interactively querying the user (e.g. active learning with embedded Human-in-the-Loop) [6]. Our motivation for this case study is to provide a proof-of-concept for setting up a model for automatic pre-selection of images from large unlabeled datasets that may be used for training ConvNets to learn the visual clues that are indicative of the quality of real estates. This work is therefore intended to serve as the basis for a more extensive follow-up study. Thus, the main incentive is to investigate how the proposed model processes complex intrinsic properties of real estate photographs, as well as which domain-specific labels are generalized well by the classifiers.

*This research was funded by the Austrian Research Promotion Agency (FFG) project 880546 "IMREA" and we are very grateful to DataScience Service GmbH for providing the data for this study.

[1]M. Despotovic and Z. Zhang are with the Kufstein University of Applied Sciences, Kufstein 6330, Tirol, Austria (miroslav.despotovic@fh-kufstein.ac.at; zedong.zhang@fh-kufstein.ac.at)

[2]E. Stumpe and M. Zeppelzauer are with the ICMT Institute of Creative Media Technologies, St. Pölten University of Applied Sciences, St. Pölten 3100, Lower Austria, Austria (estumpe@fhstp.ac.at; matthias.zeppelzauer@fhstp.ac.at)

Real estate images have different resolutions or were taken under different lighting conditions with varying distances and angles to the object. An additional challenge is that there are only a limited number of relevant labels, and it is a priori unclear which classes can even be captured from the images. The data contains noise, samples that cannot be attributed to a specific property characteristic, as well as redundant information because real estate developers in local markets often work with multiple agencies for advertising and sales.

III. APPROACH

We make the naive assumption that empirical error in the decision boundary can be minimized by exploiting the generalization capability of multiple ConvNets, provided that a large amount of training data is available. In this regard, we propose a SSL procedure as follows.

A. Iterative training

We use annotated data to iteratively fine-tune VGG16 [8] and ResNet101v2 [2] (both pre-trained on the large ImageNet dataset), starting from the initial training dataset S_i. That is, after each complete iteration, we infer labels in the unlabeled dataset $\mathcal{D}_\cup$ with fine-tuned networks N_1 and N_2 and enrich training datasets S_1 and S_2 (one set per network) with new instances. Thereby, we select randomly, at a lower threshold of 100% accuracy, 5 predictions per class and network and add them as new instances to the prior training sets. This process is performed sequentially until we obtain training sets S_1 and S_2 with 5000 instances each. The selection of 5 matches per class is deliberate to reduce the target risk due to the learner's prior knowledge [7]. The determination of false predictions in the S_1 and S_2 is carried out within the definition of experiment baselines (see IV-C).

B. Ensemble decision

We build a dataset S_{tr} consisting solely of instances of S_1 and S_2 that are predicted in concordance by both networks. The inference of the SSL model is then evaluated by fine-tuning a VGG16 with S_{tr} and testing it with an independent dataset T_1.

IV. EXPERIMENTAL SETUP

A. Data

The preprocessing of the data initially involves duplicate removal by image-wise assignment of unique hash values and calculating difference using Hamming distance. After this step, our experimental data set $\mathcal{D}_\cup$ eventually comprises 47k images. However, some redundant information remains,

as agencies often add their logos when editing photos or post-processing the image for marketing purposes.

Fig. 1. Experimental selection of real estate classes, Image source: [1]

For our study, we use a manually pre-selected ground truth set $\mathcal{D}_t$ with 12 meaningful classes from the perspective of real estate valuation. Figure 1 shows the experimental class selection. This set is then partitioned into training S_i, validation V_1 and test T_1 datasets with a ratio of 1473-375-240 instances and 12 balanced classes per set. We control our experiment by setting multiple baselines (see IV-C) with training sets S_i, S_3 and S_4 (see Table I). S_3 is a manually selected subset of S_1 where only correctly predicted labels are kept. S_4 is defined like S_3 with the exception that the incorrectly predicted labels are not excluded but manually added to the images with correctly predicted labels from S_1.

B. Setup & Training

For the training we utilize extensive data augmentation including centering, rescaling and shifting. Training parameters for both nets are learning rate of 0.001, decay of 0.001, momentum of 0.9 and a batch size of 40 for N_1 resp. 100 for N_2 . All nets were trained with cross-entropy loss and adamax optimizer [3]. A full SSL iteration was initially set to 200 epochs and successively reduced: $I_1 = 200, I_2 = 200//2, I_3 = 200//3, I_4 = 200//4, ..., I_n = 200//4$. Since we observed higher loss/accuracy variability in the earlier and later training phases, a larger number of epochs was deliberately chosen. Thus, we do not apply early stopping for regularization but select the training stage with the best performance.

C. Evaluation

We aim at answering following research questions: (1) are the individual classes sufficiently discriminative to achieve an acceptable generalization of the classifier? and (2) can the proposed experimental SSL approach achieve a comparable result to the established baselines? To measure the performance of the model, we set up multiple baselines whose performance was evaluated with the test set T_1. The lower baseline is defined as the performance of a fine-tuned VGG16 trained on initial training set S_i. The mid baseline is specified through the performance of a fine-tuned VGG16 trained on S_3. Finally, we define an upper baseline as the performance of a fine-tuned VGG16 trained on S_4.

V. RESULTS

In the Figure 2 showing Receiver Operating Characteristic (ROC) for each predicted class, a larger deviation is noticable for class 4 (map), followed by class 12 (surrounding) and class 10 (balcony/terrace). These are basically classes that do not represent interior spaces. An expected confusion can be seen between class 1 (building facade) and class 3 (building CAD). On the other hand, all classes with interiors were particularly well recognized by the classifier, indicating their discriminative visual content. However, false-positive test results point to a minor misinterpretation for classes attic and staircase.

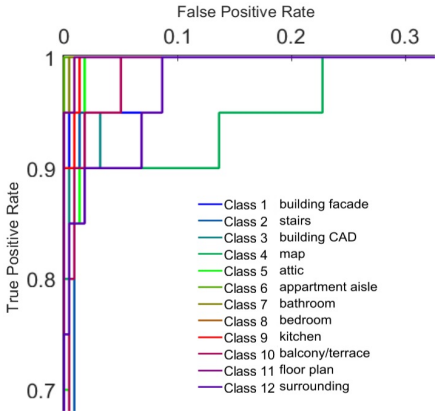


Fig. 2. ROC of individual classes.

Table I shows that the SSL model slightly underperforms lower and middle baseline, but the performance is almost consistent with the upper baseline. This is attributed to the larger proportion of false positives for classes stairs, building facade and building CAD in S_{tr} (compared to S_1 and S_2) and thus the inconsistent class balance during the training. Notably, the overall class balance in S_{tr} (intentionally not supervised) expressed by coefficient of variation CV (18%) is smaller than CV for S_3 (30.2 %) and S_4 (41.7 %).

With this study, we have gained first insights into the challenging task of enriching metadata from real estate images. We intend to build on the results of the presented approach in a more comprehensive follow-up study to gain further valuable evidence.

TABLE I
COMPARISON OF CLASSIFICATION ACCURACY (IN %) FOR SSL MODEL AND BASE MODELS.

training dataset	VGG16 lower baseline			training dataset	VGG16 mid baseline		
	sample size	validation	test		sample size	validation	test
S_1	1473	97.28	92.08	S_3	2661	96.2	91.67
training dataset	VGG16 upper baseline			training dataset	VGG16 ResNet101v2 ssl		
	sample size	validation	test		sample size	validation	test
S_4	3448	95.92	90.00	S_{tr}	3005	94.84	89.17

REFERENCES

[1] Justimmo- einfach makeln! B&G Consulting & Commerce GmbH. [Online]. Available: https://www.justimmo.at/

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[3] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.

[4] D. Koch, M. Despotovic, M. Sakeena, M. Döller, and M. Zeppelzauer, "Visual estimation of building condition with patch-level convnets," *Proceedings of the 2018 ACM Workshop on Multimedia for Real Estate Tech*, 2018.

[5] V. Muhr, M. Despotovic, D. Koch, M. Döller, and M. Zeppelzauer, "Towards automated real estate assessment from satellite images with cnns," in *Proceedings of the 10th Forum Media Technology (FMT)*, vol. 2009, 2017, pp. 14—-23.

[6] C. Persello and L. Bruzzone, "Active and semisupervised learning for the classification of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 11, pp. 6937–6956, 2014.

[7] P. Pourashraf and N. Tomuro, "Use of a large image repository to enhance domain dataset for flyer classification," in *ISVC*, 2015.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2015.

[9] Y. Zhang and R. Dong, "Impacts of street-visible greenery on housing prices: Evidence from a hedonic price model and a massive street view image dataset in beijing," *ISPRS International Journal of Geo-Information*, vol. 7, no. 3, p. 104, 2018.

Human Tracking and Pose Estimation for Subsurface Operations

Roland Perko[1], Hannes Fassold[1], Alexander Almer[1], Robert Wenighofer[2], and Peter Hofer[3]

***Abstract*— Human lives are particularly at risk in critical security situations in underground train stations compared to surface events. Due to the *closed situation* of such subsurface events, considerable obstacles to the safe and efficient evacuation of people after an attack must be taken into account. Thus, this work presents a computer vision system based on artificial intelligence that uses available surveillance cameras in the optical and the thermal spectrum to detect and track human beings, and to allow an activity classification based on a pose estimation. Those results are then transferred into a 3D common operational picture to assist subsurface operations.**

I. INTRODUCTION

Subsurface structures, like the whole subway infrastructure, are indispensable for modern societies. To ensure safety and efficient reaction to crisis, a deep understanding of the underground structure is necessary for specially trained and equipped personnel, aware of the associated risks and dangers – the so called *Subsurface Operators* [2]. In the special case of a terrorist attack available technical infrastructure can be employed to derive valuable information for those operators. Since most subsurface structures are equipped with surveillance cameras, the aim of this work is to analyse that data to assist the crisis team. From the computer vision perspective, three important queues can be derived: (1) Detection of objects of interest, in particular humans and vehicles, (2) tracking of those objects over time, and (3) activity recognition, in particular if humans are walking, standing, sitting, or lying.

With the knowledge of the coarse location and orientation of the cameras, the detections can be projected onto a map or a 3D model, which then serves as a common operational picture within a virtual reality system. To simulate the subsurface environment the test site *Zentrum am Berg* (ZaB) is chosen which allows underground research, development, training, and education at 1:1 scale [7]. An exemplary view of one of the tunnel tubes is depicted in Figure 1. This specific facility is equipped with multiple optical and thermal cameras (which are very important since critical events often occur in low or no light conditions) that serve as input for the developed computer vision and artificial intelligence system.

II. METHOD

This section reports an object detection and tracking, pose estimation for activity recognition, and the common operational picture.

[1]Roland Perko, Hannes Fassold, and Alexander Almer are with Joanneum Research, Austria {firstname.lastname}@joanneum.at

[2]Robert Wenighofer is with Montanuniversität Leoben, Austria robert.wenighofer@unileoben.ac.at

[3]Peter Hofer is with the Theresianische Militärakademie, Austria peter.hofer@bmlv.gv.at

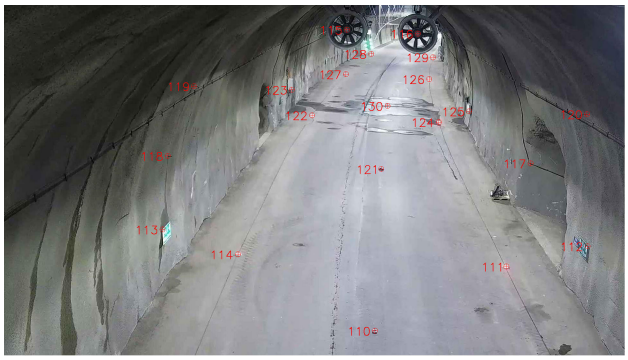


Fig. 1. Subsurface environment at ZaB with ground control points shown in red color for extrinsic camera calibration.

A. Object Detection and Tracking

For the detection and tracking of persons (and other objects), we base upon the *OmniTrack* algorithm [1]. It is real-time capable and combines a powerful deep learning based object detector (YoloV3 [8]) with high-quality optical flow methods (TV-L^1 [11]). Within this work, we updated the key components of the algorithm to more recent methods. Specifically, for the object detector component we switched from YoloV3 to the *Scaled-YoloV4* method [10]. It achieves higher accuracy by employing a cross-stage partial network and can be easily scaled to multiple resolutions. Additionally, instead of the classical TV-L^1 algorithm for optical flow we employ the recently proposed *RAFT* optical flow algorithm [9]. The RAFT optical flow method achieves high accuracy of the motion field and generalizes well to other domains (like thermal images which have a different characteristic than RGB images). Note that for both RGB and thermal input images, we use the standard Scaled-YoloV4 pretrained model, which has been trained on the MS COCO dataset [6] (consisting of RGB images). We do not fine-tune or retrain on a specific thermal image dataset. For the purpose of the project we only use the two classes humans and vehicles.

B. Human Pose Estimation

For human pose estimation, we employ the *EvoSkeleton* algorithm [5]. The method evolves a limited dataset to synthesize unseen 3D human skeletons based on a hierarchical human representation and heuristics inspired by prior knowledge. Via this special data augmentation procedure, EvoSkeleton achieves state-of-the-art accuracy on the largest public benchmark (Human3.6M [3]) and additionally generalizes well to unseen and rare poses. In order to calculate the poses (skeletons with 17 joints) for all detected persons in one frame, we proceed as follows. First, for all detected

persons the rectangular regions of interest are extracted to a list of sub-images. These sub-images are now processed in multiple batches, with the size of the batch set to 4 sub-images. The batching mechanism makes inference more efficient and ensures that the GPU memory is not exhausted. With a batch size of 4, roughly 5 GB GPU RAM are occupied. The thermal images are transferred to a different color range, which improves the performance of the pose estimation.

C. Common Operational Picture

Since all information is gathered in image geometry, it has to be transferred to the map projection of the 3D common operational picture. Therefore, the cameras are calibrated intrinsically using planar calibration random dot targets. The extrinsics are determined using ground control points (cf. the red points in Figure 1) within a least squares parameter adjustment. This calibration allows 2D information in image geometry to be intersected with an existing 3D tube model that was acquired via terrestrial laser scanning for the whole ZaB subsurface test site.

III. RESULTS

Figures 2 and 3 depict one video frame of an optical, respectively, thermal camera superimposed with the bounding boxes from the human detection and the human skeleton for each detection.

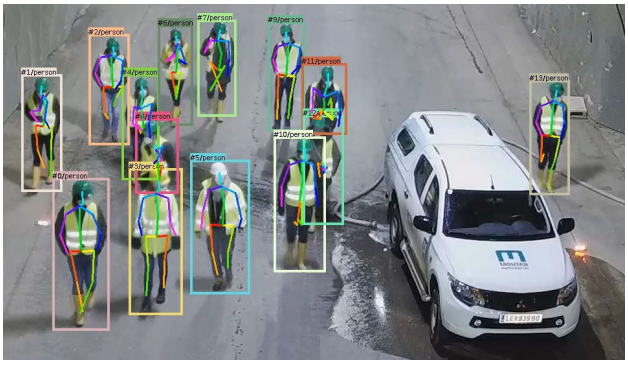


Fig. 2. Human detection and skeleton estimation for an optical image.

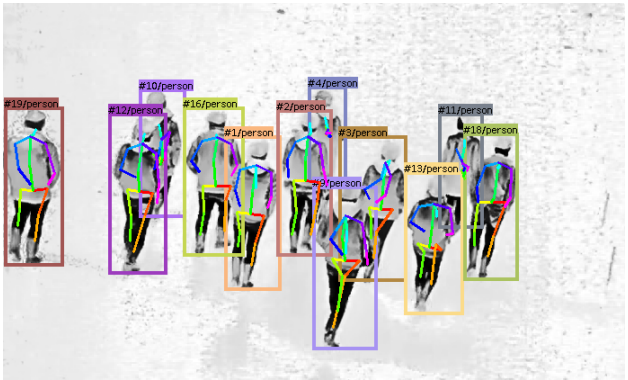


Fig. 3. Human detection and skeleton estimation for a thermal image. The color coding scheme of the thermal image was altered to improve the quality of the pose estimation.

Figure 4 depicts a screenshot of the 3D common operational picture within a virtual reality system where the subsurface operators get a simplified overview of the human detections and classifications.

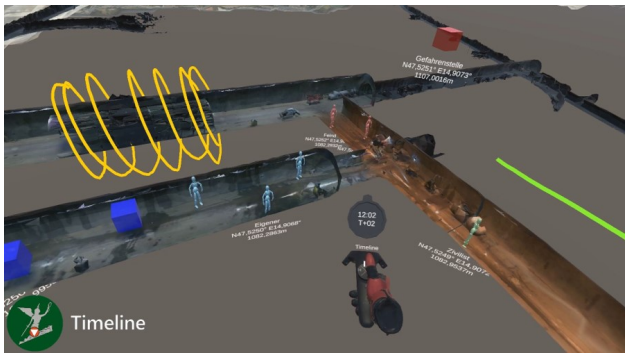


Fig. 4. The 3D common operational picture within a virtual reality system (illustration courtesy of [4]).

For RGB video, initial experiments show that both the object detection and tracking, and also the pose estimation work very well. For thermal video, the results are worse, especially for the pose estimation. This can be attributed to the *domain gap*, the fact that both methods have originally been trained on RGB image datasets and not on thermal images. Nonetheless, it seems that even on thermal video the result of the pose estimation is good enough for our task of activity classification of persons. Regarding runtime, the object detector and tracker works in real-time, whereas the pose estimation is not real-time capable. We will investigate techniques like 16-bit inference or frame subsampling in order to achieve real-time performance also for the pose estimation.

IV. CONCLUSION

A computer vision system for human tracking and pose estimation was presented, custom-tailored for subsurface operations, based on existing surveillance infrastructure. In the future, the results from the pose estimation, together with the motion information of the tracked persons, will be used for activity classification. Specifically, via the motion information a person could be classified either as stationary or moving (walking / running). Furthermore, the pose estimation information will be used to for activity recognition, in particular, whether a person is standing or lying, by analysing the person's spine orientation. Another future research focus is to preserve the privacy of people, where one option would be to use only thermal cameras.

ACKNOWLEDGMENT

The presented research activity is embedded into the project NIKE-SubMovCon #879720 within the Austrian Security Research Programme KIRAS, funded by the Austrian Research Promotion Agency (FFG).

REFERENCES

[1] H. Fassold and R. Ghermi, "OmniTrack: Real-time detection and tracking of objects, text and logos in video," in *Proc. ISM*, 2019.

[2] P. Hofer, "Coping with complexity. The development of comprehensive subsurface training standards from a military perspective," *BHM Berg-und Hüttenmännische Monatshefte*, vol. 164, no. 12, pp. 497–504, 2019.

[3] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.

[4] Laabmayr, "Subsurface operation mission tool," www.laabmayr.at/tunnel-plus/rd/somt-subsurface-operation-mission-tool/, 2021, (accessed October 18, 2021).

[5] S. Li, L. Ke, K. Pratama, Y.-W. Tai, C.-K. Tang, and K.-T. Cheng, "Cascaded deep monocular 3d human pose estimation with evolutionary training data," in *Proc. CVPR*, 2020.

[6] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *ECCV*, 2014, pp. 740–755.

[7] Montanuniversität Leoben, "Zentrum am Berg," https://www.zab.at/, 2021, (accessed October 14, 2021).

[8] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *ArXiv*, vol. abs/1804.02767, 2018.

[9] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," *ArXiv*, vol. abs/2003.12039, 2020.

[10] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Scaled-YOLOv4: Scaling cross stage partial network," *ArXiv*, vol. abs/2011.08036, 2020.

[11] A. Wedel, T. Pock, C. Zach, H. Bischof, and D. Cremers, "An improved algorithm for TV-L1 optical flow," in *Statistical and Geometrical Approaches to Visual Motion Analysis*, 2008.

Multi-Spectral Segmentation with Synthesized Data for Refuse Sorting

Harald Ganster[1] Alfred Rinnhofer[1] Georg Waltner[2] Christian Payer[2]
Heimo Gursch[3] Christian Oberwinkler[4] Reinhard Meisenbichler[4] Horst Bischof[2]

***Abstract*— Refuse sorting is a key technology to increase the recycling rate and reduce the growths of landfills worldwide. However, monitoring and parameterization of sorting facilities is still done in a mostly static fashion. This work combines multi-spectral imaging with deep learning based image recognition to monitor and dynamically optimize processes in sorting facilities. Our solution is capable of monitoring the sorting process remotely avoiding potentially harmful working conditions due to dust, bacteria, and fungal spores. Furthermore, the introduction of objective sorting performance measures enables informed decisions to improve the sorting parameters and react quicker to changes in the refuse composition.**

I. INTRODUCTION

The global refuse production is still on the increase worldwide, since the refuse output increases faster than the recycling rates [9]. The ever-changing refuse composition poses a major challenge to automated sorting in recycling application. This work presents preliminary findings of KI-Waste [5] capturing the refuse composition on conveyor belts in a refuse sorting facility. This is done by multi-spectral imaging and deep learning for semantic segmentation and object recognition on refuse streams at key points in the sorting facility.

II. RELATED WORK

Sorting facilities extract usable fractions with sorting and shredding machines connected by conveyor belts [4]. Image recognition applied to the refuse streams on these conveyor belts is capable of capturing the refuse composition since different substances have different spectral reflection characteristics. Thus, multi-spectral cameras can provide a spectral fingerprint of the material streams on the conveyor belts [12], [13]. A four-channel setup is often used consisting of RGB plus near-infra-red (NIR) cameras [3], [15], [11], [10]. In addition to these two-dimensional (2D) multi-spectral systems, a tree-dimensional (3D) acquisition can capture geometric properties useful in automatic material separation.

The resulting images of the refuse on the conveyor belt are the input for image recognition software identifying predefined refuse categories on a pixel-wise basis. Traditional image recognition techniques based on color and gradient features are typically not able to handle the large variations in appearance and shape occurring in mixed-material streams. Convolutional neural networks (CNNs) [7] have shown great performance on a variety of image recognition tasks including semantic segmentation [8], where a category label is assigned to each pixel of an image. The results of the image recognition are a good basis for predictive maintenance, optimization, automation and self-adaptation of the refuse sorting process [14].

III. IMAGE CAPTURING AND CLASSIFICATION

The high variety in substances and the challenging environmental conditions like dust, dirt, lighting, temperature, and vibrations make the image capturing challenging. We overcome these challenges by employing a line-scan-based multi-spectral system with a light-sectioning method that uses laser-line projection to determine surface profiles. It outputs high-quality four-channel multi-spectral 2D images and 3D registered image data.

The hardware setup is designed so that all capturing devices cover the same acquisition area. Nevertheless, calibration methods are required, registering the captured image data to each other. Finally, all image modalities are transformed into one common coordinate system by geometric mapping, ensuring that each pixel has a direct correspondence between geometric and spectral information.

The image classification segments each image pixel-wise into the predefined refuse categories by state-of-the-art fully-convolutional CNNs with a huge number of trainable parameters. To set these parameters in a meaningful way, CNNs need to be trained with hundreds or thousands of representative ground truth images, where each pixel is correctly annotated with its category.

Creating this ground truth manually requires an enormous labeling effort. Hence, this project uses empty belt images and images of mono-material refuse streams to effortlessly create ground truth labels as shown in Fig. 1 (top row). With this groundtruth, we can synthetically create realistic mixed-material images with known proportions and locations of refuse categories. This way we can generate unlimited amounts of annotated mixed-material images as depicted in Fig. 1 (bottom row).

IV. INITIAL PROJECT RESULTS

We train and evaluate the proposed approach within the DeepLabv3+ [2] framework. Our training consists of two steps. First, we train a binary segmentation model to distinguish belt vs. waste. In the second step, we use this model to generate groundtruth for the known mono-material images

*This work was supported by Land Steiermark within the research program „Zukunftsfonds“(No. 1330).

[1]JOANNEUM RESEARCH Forschungsgesellschaft mbH, DIGITAL, harald.ganster@joanneum.at

[2]Graz University of Technology, Institute of Computer Graphics and Vision waltner@icg.tugraz.at

[3]Know-Center GmbH hgursch@know-center.at

[4]Komptech GmbH

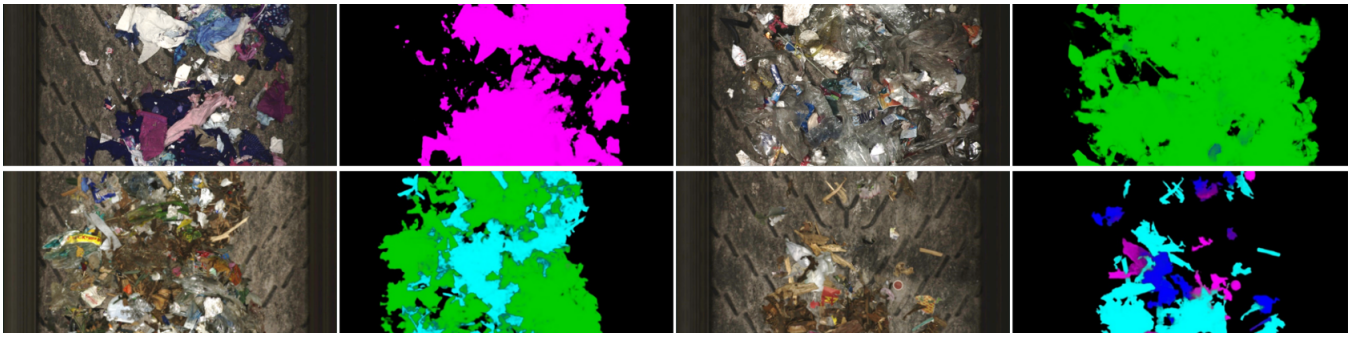


Fig. 1. Mono-material stream samples and binary segmentation belt/waste (top row) and synthetic ground-truth data with mixed stream (bottom row). All images: plastics (green), wood (turquoise), textiles (purple), paper (blue).

and use this groundtruth to train a model on synthetic mixed-material images. Manually labeling only a few waste images is sufficient for the binary network to train an initial model, which is further improved with the automatically generated labels it produces. Having a well working binary model, we use it to label the mono-material images. We then generate 130k superpixels [1] of different sizes from 492 mono-material images for our synthetic training regime, holding back 324 for testing. During training of our multi-class model, the synthetic mixed-material images are generated on-the-fly to guarantee a diverse training set. We train the network for 500k iterations with batch size 8 and Adam optimizer [6] using an initial learning rate of 0.0001 and a decay of 0.1.

As it is almost impossible to manually annotate mixed-material streams even for a trained person, we limit evaluation of the multi-class model to mono-material recordings. A great accuracy of $84-100\%$ can be observed on the refuse fractions *clothes*, *paper*, *plastic* and *wood*. Most of the fractions are very well classified except for *wood* that is partly misclassified as paper, as the confusion matrix in Fig. 2 shows. The reduction of these confusions is topic of an ongoing refinement and validation. In addition, while we cannot measure the performance due to the lack of groundtruth, we can visually observe very promising results produced by our trained CNN model also on real mixed-material streams, as shown in Fig. 3.

	belt	clothes	paper	plastic	wood
belt (24)	100.00	0.00	0.00	0.00	0.00
clothes (75)	0.00	99.88	0.04	0.02	0.06
paper (90)	0.00	1.09	97.46	1.44	0.01
plastic (90)	0.00	2.84	0.13	95.75	1.28
wood (45)	0.00	2.92	10.46	2.35	84.27

Fig. 2. Confusion matrix with pixel-wise accuracies in % for 324 test images. Apart from minor confusions between *wood* and *paper*, the performance of the CNN model is very promising.

Several properties of the visible refuse can be calculated when combining the semantic segmentation output with the 3D surface information, e. g. refuse category distribution, particle size, and the height of specific regions of the image. These properties will then be used for further analysis and refuse processing parameters adjustment.

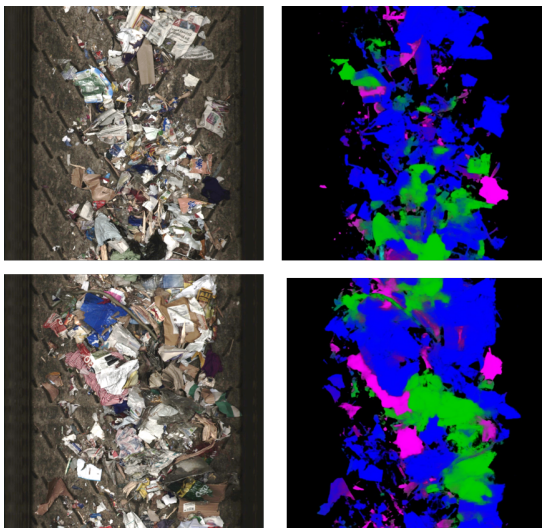


Fig. 3. Semantic segmentation results with plastics (green), wood (turquoise), textiles (purple), paper (blue).

V. Conclusion & Outlook

The project is currently in its first phase focusing on hardware design, interface definition, and data collections. The ground-truth generation strategy of using single refuse categories as starting point to synthesise realistic refuse mixtures proofed to be extremely valuable and brought a tremendous speed-up in necessary data generation, which can also have an impact on other similar deep learning applications. This data collection is the basis for all future work in the project including improvements of the camera and lighting setup, training of image recognition models, domain-specific adaption and improvements of the image recognition models, validation of the image recognition results, and all further analysis and optimizations.

Initial results of semantic segmentation and refuse classification already showed the feasibility of the approach, which will be further refined during the ongoing project and applied to other machinery on the refuse processing chain as well as to other sorting facilities in the future.

REFERENCES

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC Superpixels Compared to State-of-the-art Superpixel Methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 34, no. 11, pp. 2274–2282, 2012.

[2] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," in *Proceedings of the European Conference on Computer Vision 2018 (ECCV 2018)*, ser. Lecture Notes in Computer Science. Cham, Switzerland: Springer International Publishing, 2018, pp. 833–851.

[3] J. F. Deprez, "Hyperspectral Analysis For Precision Optical Sorting," 2017, Proceedings of the Conference on Hyperspectral Imaging in Industry (CHII 2017).

[4] S. P. Gundupalli, S. Hait, and A. Thakur, "A Review on Automated Sorting of Source-Separated Municipal Solid Waste for Recycling," *Waste Management*, vol. 60, pp. 56–74, 2017.

[5] H. Gursch, H. Ganster, A. Rinnhofer, G. Waltner, C. Payer, C. Oberwinkler, R. Meisenbichler, and R. Kern, "KI-Waste - Combining Image Recognition and Time Series Analysis in Refuse Sorting," in *Mensch und Computer 2021 - Workshopband*, C. Wienrich, P. Wintersberger, and B. Weyers, Eds. Bonn, Germany: Gesellschaft für Informatik e.V., 2021, pp. 1–4. [Online]. Available: https://dl.gi.de/handle/20.500.12116/37354

[6] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*. New York, USA: IEEE, June 2015, pp. 3431–3440.

[9] A. Minelgaitė and G. Liobikienė, "Waste Problem in European Union and its Influence on Waste management Behaviours," *Science of The Total Environment*, vol. 667, pp. 86–93, June 2019.

[10] J. Piao, Y. Chen, and H. Shin, "A New Deep Learning Based Multi-Spectral Image Fusion Method," *Entropy*, vol. 21, no. 6, pp. 1–16, June 2019.

[11] P. Prayagi, "Prism Based Multi-Sensor Technology for Multispectral Imaging Applications," 2017, Proceedings of the Conference on Hyperspectral Imaging in Industry (CHII 2017).

[12] A. Rinnhofer, "Combination of Multispectral and Multisensory Data," 2019, Fraunhofer Vision Technology Day, INNOVATIVE TECHNOLOGIES FOR INDUSTRIAL QUALITY ASSURANCE WITH IMAGE PROCESSING.

[13] M. Sackewitz, *Leitfaden zur hyperspektralen Bildverarbeitung*. Stuttgart, Germany: Fraunhofer Verlag, 2019.

[14] R. Sarc, A. Curtis, L. Kandlbauer, K. Khodier, K. E. Lorber, and R. Pomberger, "Digitalisation and Intelligent Robotics in Value Chain of Circular Economy Oriented Waste Management - A Review," *Waste Management*, vol. 95, pp. 476–492, July 2019.

[15] P. Wollmann, "Hyperspectral Imaging for Surface Inspections," 2017, Proceedings of the Conference on Hyperspectral Imaging in Industry (CHII 2017).