 DOI: 10.3217/978-3-85125-869-1-05

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

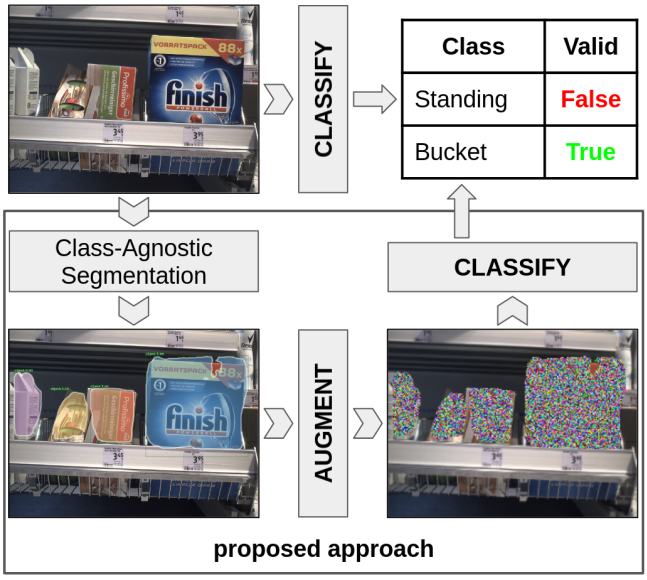

Fig. 1. Approach overview: Via class-agnostic segmentation we guide class prediction towards focusing on relevant image parts.

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

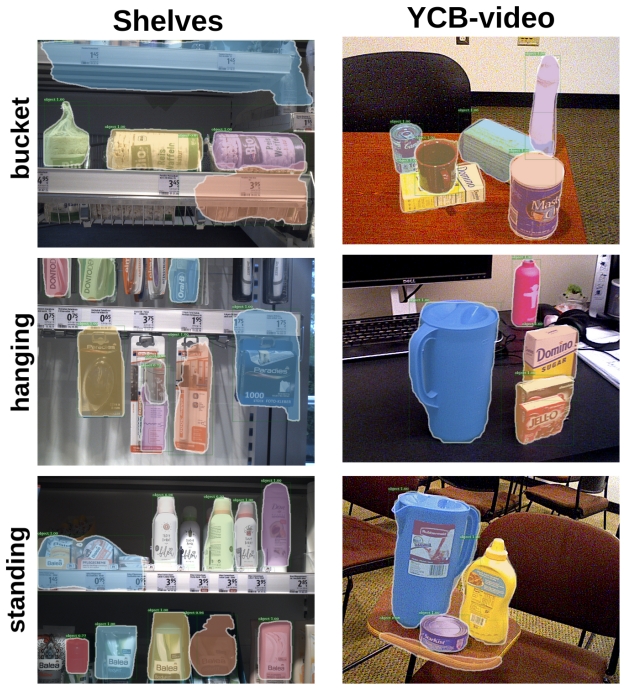

Fig. 2. Class-agnostic instance segmentation results on images of our training set for shelves (left column) and on unseen images of YCB-video [14] (right column).

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.