 DOI: 10.3217/978-3-85125-869-1-11

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.

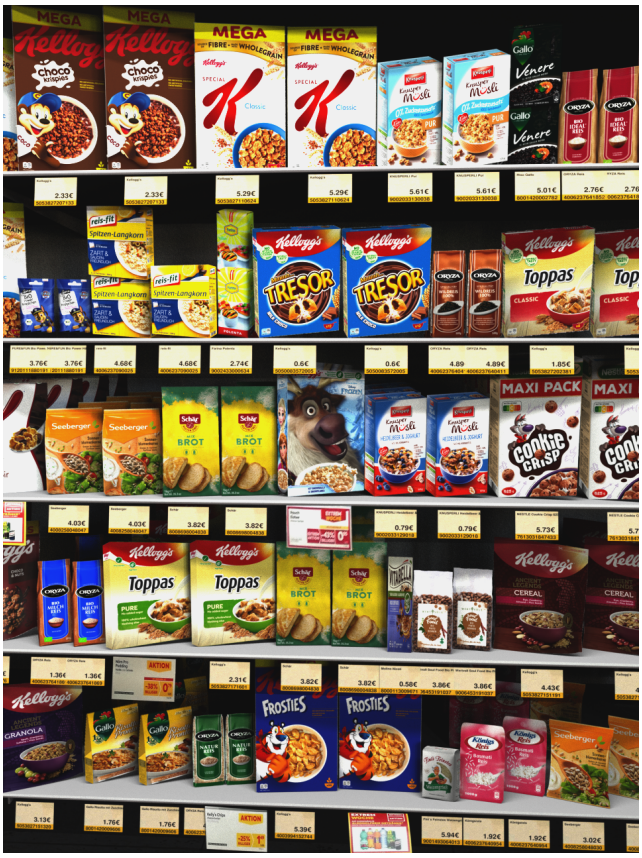

Fig. 1: Synthetic shelf image with cereal products, generated with the proposed data synthesis pipeline.

*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

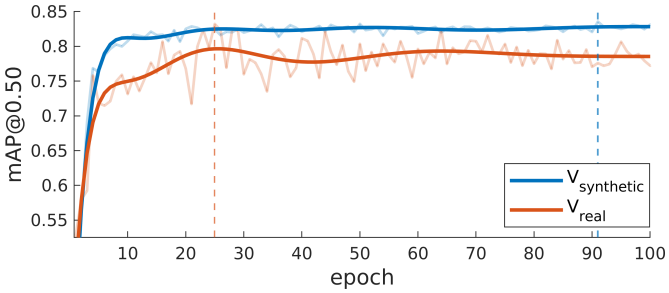

Fig. 2: Measured detection accuracy on $V_{synthetic}$ and V_{real} over 100 training epochs.

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.