CAD Models to Real-World Images: A Practical Approach to Unsupervised Domain Adaptation in Industrial Object Classification

Dennis Ritter[1], Mike Hemberger[2], Marc Hönig[3], Volker Stopp[3], Erik Rodner[4], and Kristian Hildebrand[1]

[1] Berliner Hochschule für Technik
[2] nyris GmbH
[3] topex GmbH
[4] KI-Werkstatt/FB2, University of Applied Sciences Berlin

Abstract. In this paper, we systematically analyze unsupervised domain adaptation pipelines for object classification in a challenging industrial setting. In contrast to standard natural object benchmarks existing in the field, our results highlight the most important design choices when only category-labeled CAD models are available but classification needs to be done with real-world images. Our domain adaptation pipeline achieves SoTA performance on the VisDA benchmark, but more importantly, drastically improves recognition performance on our new open industrial dataset comprised of 102 mechanical parts. We conclude with a set of guidelines that are relevant for practitioners needing to apply state-of-the-art unsupervised domain adaptation in practice. Our code is available at `https://github.com/dritter-bht/synthnet-transfer-learning`.

1 Introduction

Recognizing machine parts requires in-depth industrial domain knowledge. However, particularly in engineering, machine-specific specialists are often needed to identify components without prolonged research, making it challenging for customers of machine manufacturers to independently identify the parts of their machines. Automatic visual recognition seems therefore a straightforward solution to apply. However, complex machines typically comprise hundreds or even thousands of individual parts. Generating and labeling sufficient images of each component for training is often too costly. In contrast, companies own the computer-aided design (CAD) data of the parts, which can be rendered with any parameters and in any quantity. Consequently, our goal (Fig. 1) is to use CAD data and train a classifier with adaptation techniques from rendered 3D objects (source domain) that can be applied to real-world images (target domain).

Our proposed contribution is twofold: First, we present a comprehensive guide designed to facilitate future research in surpassing SoTA performance (MIC [8]) on the VisDA classification challenge benchmark. We analyze the performance enhancements and their impact at the different stages of our domain adaptation (DA) pipeline, providing a blueprint from a wide range of methods already present in the vast existing literature (Sect. 5). Second, we introduce a new open dataset characterized by minimal inter-class

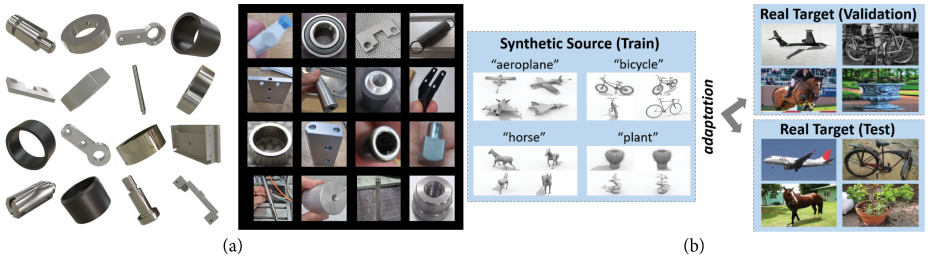

Fig. 1. (a): Our Topex-Printer dataset contains rendered and real images from 102 machine parts (Sect. 3). (b): The VisDa-2017 challenge tests UDA model performance under simulation-to-real domain shifts [22].

distances, offering a novel challenge for unsupervised domain adaptation (UDA) research (Sect. 3).

Specifically, we use publicly available models pretrained on the ImageNet22K (IN22K) dataset [1] and continue with linear probing using only source domain data to tune the classification head as initialization for further training (similar to [14]). We continue training in an unsupervised domain adaption (UDA) setting, i.e. no labels for target domain data available, applying CDAN [19] and MCC [10]. We test our approach with the VisDA-2017 image classification challenge dataset [22] and our self-made *Topex-Printer* dataset (Sect. 3) shown in Fig. 1.

2 Related Work

Adversarial training, which encourages domain-invariant image features, is a key approach in image-based DA techniques. Originally introduced in [3], it adapts the GAN concepts of [5] for DA tasks. ADDA [26] consolidates several approaches into a framework based on adversarial learning. CyCADA [7] applies CycleGAN's [30] cycle consistency for DA on image classification and semantic segmentation. CDAN [19] adds a conditional domain discriminator utilizing classifier predictions to assist the DA process. Lastly, SDAT [23] uses a *smooth task loss* to stabilize adversarial training, leading to improved generalization on the target domain.

Beyond adversarial training, discrepancy minimization methods aim to align feature representations, reducing distribution discrepancy between source and target domains. Deep Adaptation Network (DAN) [18] and JAN [20] use maximum mean discrepancy (MMD) and joint MMD for feature transfer. Contrastive Adaptation Network (CAN) [12] introduces the *Contrastive Domain Discrepancy* (CDD) metric for class-aware alignment. Sliced Wasserstein Discrepancy metric (SWD) [15] is based on the Wasserstein Distance. The *Minimum Class Confusion* (MMC) loss [10] reduces target domain cross-class confusion. Recently, Masked Image Consistency (MIC) [8] enforces prediction consistency between masked target images and complete-image pseudo-labels. Kumar et al. [14] suggest an optimized transfer learning scheme that initially updates the classification head, then fine-tunes all parameters—proves to be particularly effective for large distribution

shifts in out-of-distribution datasets by preserving pretrained features. Our work adopts this approach, combining CDAN [19] and MCC [10] for UDA. While many methods rely on CNNs, recent studies [29,13] show that Vision Transformer (ViT) [2] models surpass these. In addition, the benchmark ranking for CNNs does not extend to Transformer models, although pretraining significantly improves domain transfer [13]. For a comprehensive survey of transfer learning, encompassing pretraining and adaptation techniques, refer to [9].

We utilize the VisDA-2017 image classification dataset, comprising three subsets: a training set of 150k rendered 2D images from 1,907 3D models, a validation set of 174k real photos from MS COCO [16], and a test set of 72k real images from Youtube-boundingboxes [24]. Each image is categorized into one of twelve classes. However, as shown later, performance on this dataset already saturates and therefore a novel benchmark is required.

3 A New Domain Adaptation Benchmark: Topex-Printer

We introduce a challenging dataset for identifying machine parts from real photos, featuring images of 102 parts from a labeling machine. This dataset was developed with the complexity of real-world scenarios in mind and highlights the complexity of distinguishing between closely related classes, providing an opportunity to improve domain adaption methods. The dataset includes 3,264 CAD-rendered images (32 per part) and 6,146 real images (6 to 137 per part) for UDA and testing. Rendered images were produced using a Blender-based pipeline with environment maps, lights, and virtual cameras arranged to ensure varied mesh orientations. We also use material metadata and apply one of 21 texture materials to the objects. We render all images at 512^2 pixels. Some examples of our rendered images can be seen on the left side of Fig. 1 (a). The real photo set consists of raw images captured under varying conditions using different cameras, including varied lighting, backgrounds, and environmental factors. More examples are available in the supplementary material. The dataset is publicly available at `https://huggingface.co/datasets/ritterdennis/topex-printer/resolve/main/topex-printer.zip`.

4 Our adaptation pipeline

We reviewed existing research, analyzing two prevalent stages of DA training. This led to our empirically-backed approach that yielded robust results on the Topex-Printer and VisDA datasets, achieving 93.47% accuracy on the target domain for the latter, which exceeds the accuracy reported in [8]. The steps comprise the following:

1. **Adapting pretrained models to rendered images:**
   (a) We start from pretrained models and train a new classification head with source domain data (see [14,13]). For this, we freeze layers, exchange the class head to the necessary number of classes and tune the class head with source data only (CH).

   (b) We executed a fine-tuning across all layers and a hyperparameter search (optimizer, scheduler, learning rate, augmentations) for our DA experiments on source domain data only (FT).
2. **Adapting to real-world images with UDA:**
   (a) We use the best parameters from experiments training only with source domain data for our UDA experiments and start training from the checkpoint with the tuned classification head.
   (b) We conduct studies on our two datasets with the methods CDAN, MCC, and CDAN-MCC combined and analyze the effect of all our parameters in Sect. 5.

While these are standard procedures in DA, we lay out the most important aspects for the single steps in the next sections.

4.1 Adapting pretrained models to rendered images

We conduct transfer learning on various models (ViT [2], Swinv2 [17] and DeiT [25], please refer to the supplementary material for version details), pretrained on IN22k, using only source domain data for training and identical training procedures and configurations. This approach allows us to establish a suitable baseline and determine appropriate training parameters. First, we load the pretrained model and replace the linear classification head with one that matches the number of classes in our dataset (12 outputs for VisDa-2017 [22], 102 outputs for Topex-Printer). We perform three different training schemes: training the classification head only (CH), fine-tuning the full model (FT), and a combination of CH and FT, tuning the classification head first and continuing with full fine-tuning (CH-FT) inspired by [14].

1. For CH, we freeze all layers but the classification head and train for 20 epochs using SGD with learning rates [10.0, 1e-01, 1e-03], momentum 0.9, no weight decay, no learning rate scheduler, and no warmup.
2. For FT, we do not freeze any layers and train for 20 epochs using AdamW optimizer with learning rates [1e-01, 1e-03, 1e-05], weight decay 0.01, cosine annealing learning rate scheduler [21] without restarts, and two warmup epochs (10% of total epochs).
3. For CH-FT, we use the best-performing CH training run based on the test set's top-1 accuracy and continue fine-tuning the whole model from the best validation checkpoint using parameters of the best-performing FT run for another 20 epochs (so 40 epochs total training after pretraining).

For both datasets, VisDa-2017 and Topex-Printer, we use a batch size of 32 and two different data augmentation setups. For all runs, we use random resized crops with relative scale range (0.7, 1.0), random horizontal flip, random color jitter with parameters (brightness=0.3, contrast=0.3, saturation=0.3, hue=0.3), random grayscale, and normalize the final tensor using standard deviation [0.5, 0.5, 0.5] and mean [0.5, 0.5, 0.5]. We further replace random color jitter and random grayscale by AugMix [6] with default parameters.

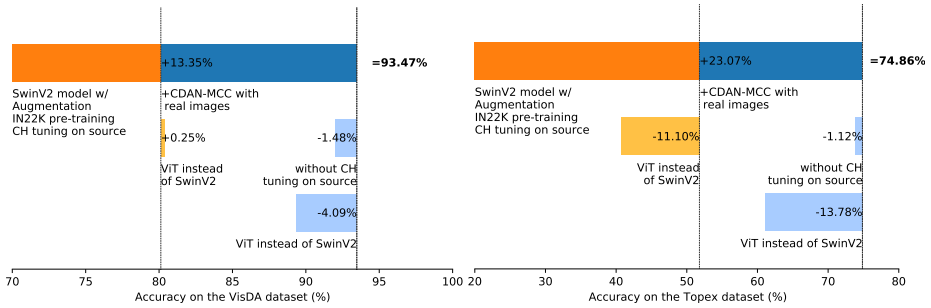

Fig. 2. Results of our DA pipeline for the (Left): VisDA and our (Right) Topex dataset. Blue bars highlight results obtained using UDA with additional target images.

Table 1. Accuracy in % on VisDA-2017 target domain (real images)

ViT model						Swin model	
TVT [29]	CDTRANS [28]	SDAT [23]	MIC [8]	Ours (w/o UDA)	Ours	Ours (w/o UDA)	Ours
83.9	88.4	89.8	92.8	80.54	89.38	80.12	**93.47**

4.2 Adapting to real-world images with unsupervised adaptation

Upon completion of the first stage, we proceed with further experiments in an UDA setting. For these, we solely employ the SwinV2 [17] and ViT [2] model architectures, as these demonstrated superior performance (see supplementary material table 3 for details). We start with the optimal classification head (CH) checkpoint from our experiments described in section 4.1 and keep the training parameters consistent with the best-performing fine-tuning (FT) run for each model. We execute six UDA runs each for the ViT [2] and SwinV2 [17] models: 20 training epochs, 32 batch size, AdamW optimizer with a 1e-05 learning rate, 1e-02 weight decay, and a cosine annealing learning rate scheduler without restarts and a two-epoch warmup (details in supplementary material). Image augmentations - random resized crop, horizontal flip, and AugMix [6] - are utilized as described in Sect. 4.1. Essentially, we replicate the process executed for source-domain-only CH-FT training runs, while concurrently incorporating UDA techniques — namely, CDAN [19] and MCC [10]. Following the findings of [13] that CDAN [19] outperforms even newer DA techniques using modern architectures (Vit-L, ConvNext-XL), we decide to use the Transfer Learning Library (tllib) [9,11] implementations of CDAN (hidden size 1024) and MCC [10] (temperature 1.0) DA methods and also combine both.

5 Evaluation

Our experiments are always based on measuring the mean class-wise accuracy in the target domain, *i.e.* the real-world images.

Results on VisDA-2017 Dataset Our first evaluation is done on the standard domain adaptation benchmark VisDA-2017 [22], where we are able to achieve SoTA performance as highlighted in Tab. 1. One can see, that our ViT training outperforms TVT [29] and achieves competitive results compared to CDTRANS [28] and SDAT [23] but does not reach the performance of MIC [8] when the same ViT architecture is used (please refer to supplementary material table 7 for details). However, our pipeline with the SwinV2 architecture slightly outperforms the current state of the art by 0.68% accuracy.
Most importantly for us and the paper, we analyzed the contribution of each part of our pipeline in Fig. 2 (left). In this figure, the results of several ablations have been visualized with blueish bars referring to results achieved with additional target images through UDA techniques. The results reveal several aspects:

1. Unsupervised domain adaptation is important to adapt to real-world images: Our best models with source data only, achieve around 80% accuracy, but with CDAN [19] and MCC [10] as combined UDA techniques, we are able to outperform all other approaches on this dataset.
2. It is beneficial and fast and easy to use class head (CH) tuning on the source data before applying UDA techniques to prevent feature distortion [14]: This can be seen in the -1.48% drop in performance without CH tuning.
3. Using the right model architecture is crucial for UDA: Our ViT models after UDA achieve less than 90% accuracy (drop of 4.09%). This difference in performance is insignificant before UDA.
4. Our SoTA performance was achieved after only 3 training epochs of fine-tuning from the pretrained checkpoint on a single Nvidia Tesla V100 PCIE 32GB GPU (CH-checkpoint after 1 epoch + 2 Epochs UDA with CDAN+MCC). However, the number of training epochs and training stability varies between our runs but almost all experiments achieve the best validation accuracy after just a few epochs of training.

Further experimental results are given in the supplementary material of the paper and reveal the following additional aspects:

1. CDAN+MCC in combination outperforms CDAN and MCC individually in most cases (see supp. table 5 and table 6).
2. Given the ConvNextV2 [27]-based runs' modest performance—12.42% and 19.82% for source-data-only experiments, we suspend further experiments with this architecture. (see supp. table 1)

Results on the Topex-Printer Dataset
The high accuracies on VisDA-2017 [22] in general and the marginal improvements achieved on this dataset in the last years, suggest the use of a more challenging dataset to benchmark domain adaptation pipelines. Therefore, we developed and assembled the Topex-Printer dataset (Sect. 3). The results on the dataset are given in Fig. 2 (Right) and similar conclusions compared to the previous section can be drawn:

1. Unsupervised domain adaptation is even more important on this dataset: with a 23.07% gain in performance, the domain gap between the rendered images and the real-world images is likely larger compared to VisDA-2017.

2. It is again reasonable to do CH tuning before UDA. Surprisingly, SwinV2 setups using CDAN [19] or MCC [10] alone do not benefit from using a tuned classification head but instead perform worse than just using the pretrained checkpoint from Huggingface (see supplementary material table 6 for these results). However, when using CDAN and MCC combined starting from the tuned classification head, the final model performs 1.12% better. For the ViT runs on the other hand, the CH initialized runs outperform runs without classification head tuning significantly.
3. The Swin-V2 model shows a remarkable performance compared to the ViT model with a performance gain of $+11.10\%$ before UDA and $+13.78\%$ after UDA.

6 Conclusion

We propose a practical approach for an image classifier in a DA setting using rendered images from 3D objects as the source domain and real images as the target domain. We conducted several experiments performing transfer learning with source data only to set a strong baseline for follow-up UDA training using the VisDA-2017 image classification challenge dataset and our newly proposed Topex-Printer dataset with more than 100 categories. In our DA experiments, we outperformed the current state-of-the-art [8] by achieving a mean accuracy of 93.47% on the VisDA-2017 dataset and 74.86% on the Topex-Printer dataset. One goal in future work is to adapt our framework to object detection scenarios [4].

Acknowledgements: This work was funded by the German Federal Ministry of Education and Research (BMBF) through their support of the project SynthNet, a part of the KMU-Innovativ initiative (project code: 01IS21002C), the KI-Werkstatt project at the University of Applied Sciences Berlin (part of the Forschung an Fachhochschulen program (project code: 13FH028KI1) as well as project TAHAI (funded by IFAF Berlin).

References

1. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. CVPR pp. 248–255 (2009)
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
3. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. The journal of machine learning research **17**(1), 2096–2030 (2016)
4. Goehring, D., Hoffman, J., Rodner, E., Saenko, K., Darrell, T.: Interactive adaptation of real-time object detectors. In: 2014 IEEE international conference on robotics and automation (ICRA). pp. 1282–1289. IEEE (2014)
5. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: NeurIPS (2014)
6. Hendrycks, D., Mu, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshminarayanan, B.: Augmix: A simple data processing method to improve robustness and uncertainty. In: ICLR (2019)
7. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A.A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation

8. Hoyer, L., Dai, D., Wang, H., Van Gool, L.: Mic: Masked image consistency for context-enhanced domain adaptation. In: CVPR. pp. 11721–11732 (2023)
9. Jiang, J., Shu, Y., Wang, J., Long, M.: Transferability in deep learning: A survey. ArXiv **abs/2201.05867** (2022)
10. Jin, Y., Wang, X., Long, M., Wang, J.: Minimum class confusion for versatile domain adaptation. In: ECCV (2019)
11. Junguang Jiang, Baixu Chen, B.F.M.L.: Transfer-learning-library. `https://github.com/thuml/Transfer-Learning-Library` (2020)
12. Kang, G., Jiang, L., Yang, Y., Hauptmann, A.: Contrastive adaptation network for unsupervised domain adaptation. CVPR pp. 4888–4897 (2019)
13. Kim, D., Wang, K., Sclaroff, S., Saenko, K.: A broad study of pre-training for domain generalization and adaptation. In: ECCV (2022)
14. Kumar, A., Raghunathan, A., Jones, R.M., Ma, T., Liang, P.: Fine-tuning can distort pretrained features and underperform out-of-distribution. In: ICLR (2022)
15. Lee, C.Y., Batra, T., Baig, M.H., Ulbricht, D.: Sliced wasserstein discrepancy for unsupervised domain adaptation. CVPR pp. 10277–10287 (2019)
16. Lin, T.Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
17. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., Guo, B.: Swin transformer v2: Scaling up capacity and resolution. CVPR pp. 11999–12009 (2021)
18. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: ICML. pp. 97–105. PMLR (2015)
19. Long, M., Cao, Z., Wang, J., Jordan, M.I.: Conditional adversarial domain adaptation. In: NeurIPS (2017)
20. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Deep transfer learning with joint adaptation networks. In: ICML (2016)
21. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv: Learning (2016)
22. Peng, X., Usman, B., Kaushik, N., Wang, D., Hoffman, J., Saenko, K.: Visda: A synthetic-to-real benchmark for visual domain adaptation. In: CVPR-W. pp. 2021–2026 (2018)
23. Rangwani, H., Aithal, S.K., Mishra, M., Jain, A., Babu, R.V.: A closer look at smoothness in domain adversarial training. In: ICML (2022)
24. Real, E., Shlens, J., Mazzocchi, S., Pan, X., Vanhoucke, V.: Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. CVPR pp. 7464–7473 (2017)
25. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jegou, H.: Training data-efficient image transformers amp; distillation through attention. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. vol. 139, pp. 10347–10357. PMLR (18–24 Jul 2021)
26. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. CVPR pp. 2962–2971 (2017)
27. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: Convnext v2: Co-designing and scaling convnets with masked autoencoders. ArXiv **abs/2301.00808** (2023)
28. Xu, T., Chen, W., Pichao, W., Wang, F., Li, H., Jin, R.: Cdtrans: Cross-domain transformer for unsupervised domain adaptation. In: ICLR (2021)
29. Yang, J., Liu, J., Xu, N., Huang, J.: Tvt: Transferable vision transformer for unsupervised domain adaptation. In: WACV. pp. 520–530 (2021)
30. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (Oct 2017)