Learnable Data Augmentation for One-Shot Unsupervised Domain Adaptation (Supplementary)

Julio Ivan Davila Carrazco[1,3]
julio.davila@iit.it
Pietro Morerio[1]
pietro.morerio@iit.it
Alessio Del Bue[1]
alessio.delbue@iit.it
Vittorio Murino[1,2,4]
vittorio.murino@iit.it

[1] Pattern Analysis and Computer Vision (PAVIS)
Italian Institute of Technology
Genoa, Italy

[2] Department of Computer Science and Technology, Bioengineering, Robotics and Systems Engineering
University of Genova
Genoa, Italy

[3] Department of Marine, Electrical, Electronic and Telecommunications Engineering Robotics and Systems Engineering
University of Genova
Genoa, Italy

[4] Department of Computer Science
University of Verona
Verona, Italy

1 Introduction

In this supplementary material, we expand the information presented for our proposed method LearnAug-UDA. In section 2, we describe the network configuration for our Augmentation Module (AUM). In section 3, we present a qualitative comparison of the augmented samples synthesized by our AUM and the baselines. In section 4, we expand the results presented for VisDA [4].

2 Encoder-Decoder description

Our proposed approach employs augmented samples that display perceptual similarities with the Target domain. These augmented samples are generated via an Augmentation module (AUM) which exploits style-transfer techniques to learn. We present two distinct versions of AUM, both of them based on an Encoder-Decoder architecture. The first version, the Shared

Encoder (SE), consists of one encoder and one decoder architecture where the conditioning is done in the bottleneck via mixup [6]. The second version, the Disentangled Enconder (DE), consists of two encoders, one bottleneck module which mixes the embeddings from the encoders, and one decoder which synthesizes the augmented sample. In both versions, i.e. SE and DE, we make use of an encoder based on UNIT's encoder [2]. For the decoder network, we based our architecture on UNIT's generator. Unlike UNIT, we change the deconvolutional layers of the decoder to an upsampling plus convolutional layer to minimize the Checkerboard artifacts on the augmented samples. Finally, the Bottleneck module is a convolutional block similar to the one used by the encoder. In Table 1, we present the network architecture for the encoder, the decoder, and the bottleneck networks.

Table 1: Network architecture for the Augmentation Module. The Shared Encoder and the Disentangled Encoders shared the same configurations for their respectives encoders.

Layer	**Encoder**
1	Conv (channels=64, kernel size=7, stride=2), Leaky ReLU
2	Conv (channels=128, kernel size=4, stride=2), Leaky ReLU
3	Conv (channels=256, kernel size=4, stride=2), Leaky ReLU
4	Residual block (channels=256, kernel size=3, stride=1)
5	Residual block (channels=256, kernel size=3, stride=1)
6	Residual block (channels=256, kernel size=3, stride=1)
7	Residual block (channels=256, kernel size=3, stride=1)
Layer	**Decoder**
1	Residual block (channels=256, kernel size=3, stride=1)
2	Residual block (channels=256, kernel size=3, stride=1)
3	Residual block (channels=256, kernel size=3, stride=1)
4	Residual block (channels=256, kernel size=3, stride=1)
5	Upsampling (Bilinear), Conv (channels=128, kernel size=3, stride=1), Leaky ReLU
6	Upsampling (Bilinear), Conv (channels=128, kernel size=3, stride=1), Leaky ReLU
7	Conv (channels=3, kernel size=3, stride=1), Sigmoid
Layer	**Bottleneck**
1	Conv (channels=256, kernel size=7, stride=1), ReLU

3 Qualitative comparisons

In this section, we present a comparison between the diverse augmented samples generated by our method and the baselines, i.e. ASM[3], TeachAugment[5], and TOS-UDA[1]. To facilitate a comprehensive comparison, all the methods were trained using the same target samples except for TeachAugment, i.e. TeachAugment does not requires target data.

3.1 DomainNet (1 Target)

In Figure 1, we illustrate a set of augmented samples synthesized by our proposed approach and the selected baselines. The augmeted samples are synthesized for the DA task of Sketch to Painting of DomainNet. By choosing this DA task, we are able to display the range of the possible augmentations that each methods is capable of. The augmented samples of TOS-UDA and TeachAugment are not capable of properly represent the color spectrum of the target image as they only work with fixed transformations. For ASM, its augmented samples

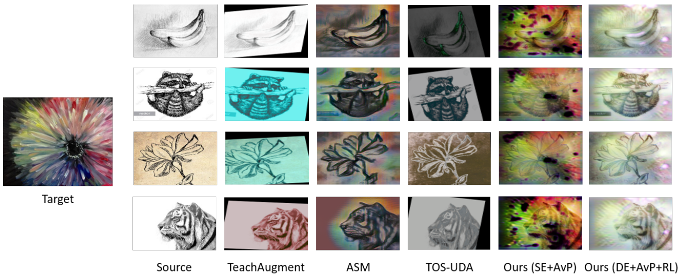

Figure 1: Qualitative comparison between our proposed approach and the selected baselines. (SE) refers to Shared encoder, while (DE) represents the Disentangled encoders. (AvgP) indicates the use of average pooling by the Style Alignment module, and (RL) specifies a model trained with the reconstruction loss.

display a perceptual similarity closer to target. However, ASM utilizes a pretrained module (RAIN) on WikiArts which results in an advantage when evaluating this specific DA task (Sketch to Painting). ASM may not have the same results for other domains. Furthermore, our augmented samples are generated by the Augmentation module which does not require pretraining to synthesize augmented samples with high perceptual similarity to the target.

3.2 Method ablations

In Figure 2, we present different augmented samples that were synthesized using different ablations of the Augmentation module (AUM). For this comparison, we trained our proposed method using three target samples (see Fig. 2 Target). These augmented samples are synthesized for the DA task of Painting to Real of DomainNet. In Table 2, we present the reported accuracies for this specific DA task to allow a better comparison of the augmented samples. Now, the presented images clearly demonstrate that applying the average pooling operation helps to smooth out hard details that are transferred from the target samples. Furthermore, the Disentangled encoders (DE) are capable of synthesizing images with less artifacts than the Shared encoder (SE), i.e. the augmented samples are less noisy therefore it obtains a better performance. Finally, introducing the reconstruction loss (RL) into the process allows the AUM to disentangle better content and style. Thus, the style encoder is capable of transferring better the characteristics of the Target domain.

Table 2: Reported accuracies for the DA task of Painting to Real for DomainNet. (SE) refers to the Shared encoder, while (DE) is the Disentangled encoders. (AvgP) indicates the use of average pooling by the Style Alignment module, and (RL) specifies a model trained with the reconstruction loss.

	SE	**SE+AvP**	**DE**	**DE+AvP**	**DE+AvP+RL**
Accurary	64.32 ± 4.42	66.21 ± 0.66	66.53 ± 1.70	69.11 ± 0.61	69.59 ± 0.41

4 VisDA results

In Table 3, we present the results for VisDA [4]. The results are obtained after performing five experiments for each of the methods. The target for each experiment was selected randomly. We present mean accuracy over all the class and their corresponding standard

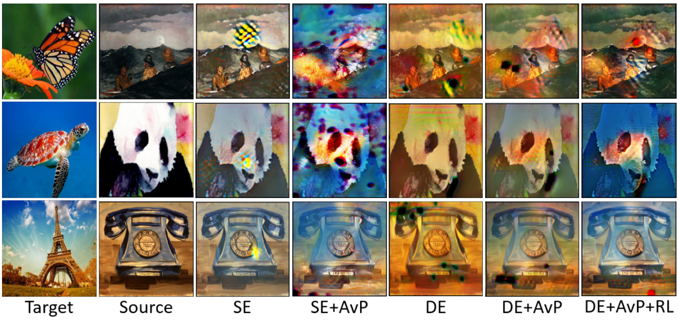

Figure 2: Qualitative comparison between different ablations of our proposed approach.

deviations. The results indicates that VisDA is a more challenging DA benchmark. The presence of large standard deviation values, particularly in certain classes, suggests that the quality of the selected target has a profound effect on the synthesized samples. However, upon observing the mean accuracy and its standard deviation, we can conclude that the proposed method consistently performs well.

Table 3: Classification accuracy of our proposed method on VisDA. For Few-shot, three target samples are used. (SE) refers to the Shared encoder, while (DE) is the Disentangled encoders. (RL) specifies a model trained with the reconstruction loss.

Method	#.T.	Aeroplane	Bicycle	Bus	Car	Horse	Knife	Motorcycle	Person	Plant	Skateboard	Train	Truck	Mean
Source only	-	**68.86** ± 8.41	3.24 ± 0.96	46.05 ± 8.73	**97.61** ± 1.15	30.48 ± 10.91	8.08 ± 3.44	50.69 ± 8.98	5.90 ± 3.66	72.14 ± 18.22	16.97 ± 4.42	62.21 ± 11.79	14.84 ± 5.14	39.76 ± 5.38
TeachAugm [■]	-	26.47 ± 3.18	0.35 ± 0.24	39.49 ± 11.79	40.38 ± 8.59	1.28 ± 0.59	1.21 ± 0.62	31.76 ± 8.75	0.40 ± 0.27	39.36 ± 11.45	9.67 ± 1.72	55.69 ± 17.22	10.45 ± 7.94	21.38 ± 1.49
ASM[■]	1	62.49 ± 9.51	**25.17** ± 7.22	**81.61** ± 4.38	77.23 ± 5.20	47.72 ± 10.28	11.84 ± 3.74	39.51 ± 12.10	5.68 ± 1.36	**83.93** ± 7.87	30.07 ± 7.08	48.77 ± 11.12	**31.49** ± 7.37	45.46 ± 1.24
TOS-UDA[■]	1	15.05 ± 12.49	0.01 ± 0.02	13.96 ± 15.88	17.31 ± 17.19	2.47 ± 3.80	20.38 ± 34.03	0.53 ± 0.29	1.34 ± 1.63	11.46 ± 15.11	7.15 ± 7.84	20.51 ± 17.48	5.32 ± 6.46	9.63 ± 6.46
Ours (DE+RL)	1	59.90 ± 6.54	12.77 ± 3.61	71.99 ± 10.45	91.46 ± 3.02	48.44 ± 5.83	**23.70** ± 7.29	**59.88** ± 5.69	11.56 ± 4.55	76.38 ± 5.99	**40.22** ± 2.00	63.19 ± 8.82	24.26 ± 4.87	48.64 ± 2.56
TOS-UDA[■]	3	21.92 ± 18.24	1.02 ± 1.76	19.66 ± 11.76	11.56 ± 20.25	7.32 ± 11.27	7.60 ± 14.49	5.15 ± 7.04	2.24 ± 4.14	11.67 ± 2.66	11.29 ± 4.05	17.90 ± 21.06	5.74 ± 5.74	10.26 ± 1.61
Ours (DE+RL)	3	62.21 ± 9.09	10.68 ± 3.02	68.38 ± 4.20	90.93 ± 3.06	**53.88** ± 3.85	22.99 ± 2.79	58.91 ± 7.16	**12.66** ± 2.85	70.14 ± 4.15	39.49 ± 4.28	**67.24** ± 2.46	27.95 ± 3.48	**48.79** ± 0.50

References

[1] Julio Ivan Davila Carrazco, Suvarna Kishorkumar Kadam, Pietro Morerio, Alessio Del Bue, and Vittorio Murino. Target-driven one-shot unsupervised domain adaptation. In *International Conference on Image Analysis and Processing*, pages 87–99. Springer, 2023.

[2] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30, 2017.

[3] Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Adversarial style mining for one-shot unsupervised domain adaptation. *Advances in neural information processing systems*, 33:20612–20623, 2020.

[4] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

[5] Teppei Suzuki. Teachaugment: Data augmentation optimization using teacher knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10904–10914, June 2022.

[6] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.