Spectral Unsupervised Domain Adaptation for Visual Recognition Supplemental Materials

A. Experiment

A.1. Datasets

Datasets for Object Detection: We consider two adaptation tasks Cityscapes [6] $\rightarrow$ Foggy Cityscapes [19] and PASCAL VOC [7] $\rightarrow$ Clipart1k [10]. Cityscapes has 2975 training images and 500 validation images, where the bounding boxes are generated from pixel-wise annotations as in [5, 18]. Foggy Cityscapes is derived from Cityscapes by adding simulated fog. We adopt all training images of Cityscapes and Foggy cityscapes as source domain and target domain, and evaluate on all validation images of Foggy cityscapes. PASCAL VOC [7] is collected from real world, while Clipart1k [10] is an artistic dataset from CMPlaces [3] and two image search engines. We adopt PASCAL VOC 2007 (with $2,501$ training images and $2,510$ validation images) and PASCAL VOC 2012 (with $5,717$ training images and $5,823$ validation images) as source domain, $1,000$ images in Clipart1k as target domain and half of Clipart1k as validation dataset as in [10, 11, 18].

Datasets for Image Classification: We consider two adaptation tasks VisDA17 [13] and Office-31 [17]. VisDA17 has $152,409$ synthetic image and $55,400$ real images with 12 categories in common. We consider the synthetic$\rightarrow$real here. Office-31 has images of 31 classes from Amazon (A), Webcam (W) and DSLR (D) which have 2817, 795 and 498 images, respectively. Following [17, 24], we study six adaptation tasks: A$\rightarrow$W, D$\rightarrow$W, W$\rightarrow$D, A$\rightarrow$D, D$\rightarrow$A, and W$\rightarrow$A.

Datasets for Semantic Segmentation: We consider two synthetic-to-real tasks including GTA5 [15] $\rightarrow$ Cityscapes [6] and SYNTHIA [16] $\rightarrow$ Cityscapes. Cityscapes here has 30 categories with pixel-wise annotations. GTA5 has $24,966$ synthetic images and shares 19 categories with Cityscapes. For SYNTHIA, we use 'SYNTHIA-RAND-CITYSCAPES' which contains $9,400$ synthetic images and shares 16 categories with Cityscapes. For the two tasks, we adopt the 2975 training images in Cityscapes as target domain and evaluate on the 500 validation images in Cityscapes.

A.2. Implementation Details

Object Detection: For Cityscapes$\rightarrow$ Foggy Cityscapes, we adopt Faster R-CNN [14] and deformable-DETR [22] as detection networks and ResNet-50 [8] as backbone as in [2, 22]. For deformable-DETR, we adopt SGD optimizer [1] with a momentum 0.9 and a weight decay $1e-4$. The initial learning rate is $2e-4$. For Faster R-CNN, we use SGD optimizer [1] with a momentum 0.9 and a weight decay $5e-4$. The initial learning rate is 0.001. For PASCAL VOC $\rightarrow$ Clipart1k, we adopt Faster R-CNN with ResNet-101 [8] as the detection network as in [10, 18]. We use SGD optimizer [1] with a momentum 0.9, a weight decay 0.0001, and an initial learning rate 0.001.

Image Classification: Following [17, 24], we use ResNet-101 and ResNet-50 [8] as backbones for the tasks VisDA17 and Office-31, respectively. We adopt SGD optimizer [1] with a momentum 0.9 and a weight decay $5e-4$. The initial learning rate is $1e-3$.

Semantic Segmentation: We use DeepLab-V2 [4] with ResNet-101 [8] as the segmentation network as in [20, 23]. We use SGD optimizer [1] with a momentum 0.9 and a weight decay $1e-4$. The initial learning rate is $2.5e-4$ and decayed by a polynomial policy of power 0.9 [4].

For all visual recognition tasks, we set the number of FCs N at 32. The weight factors λ_c and λ_s (in Eq.8 in main text) are fixed at 0.1.

B. Discussion

B.1. Number of STs

We studied the effect of the number of STs by increasing the number of ST from 1 to 4 over the UDA-based object detection task Cityscapes $\rightarrow$ Foggy cityscapes. Since the module MSL can not work with one ST, we remove MSL module in all experiments for a fair comparison. As shown in Table 1, we can observe that employing two STs performs clearly better than employing one ST. However, when the number of ST continues to increase, the performance of model doesn't improve further, demonstrating that the domain adaptation saturates with more STs. Meanwhile, more STs will complicate the network design and introduce more parameters.

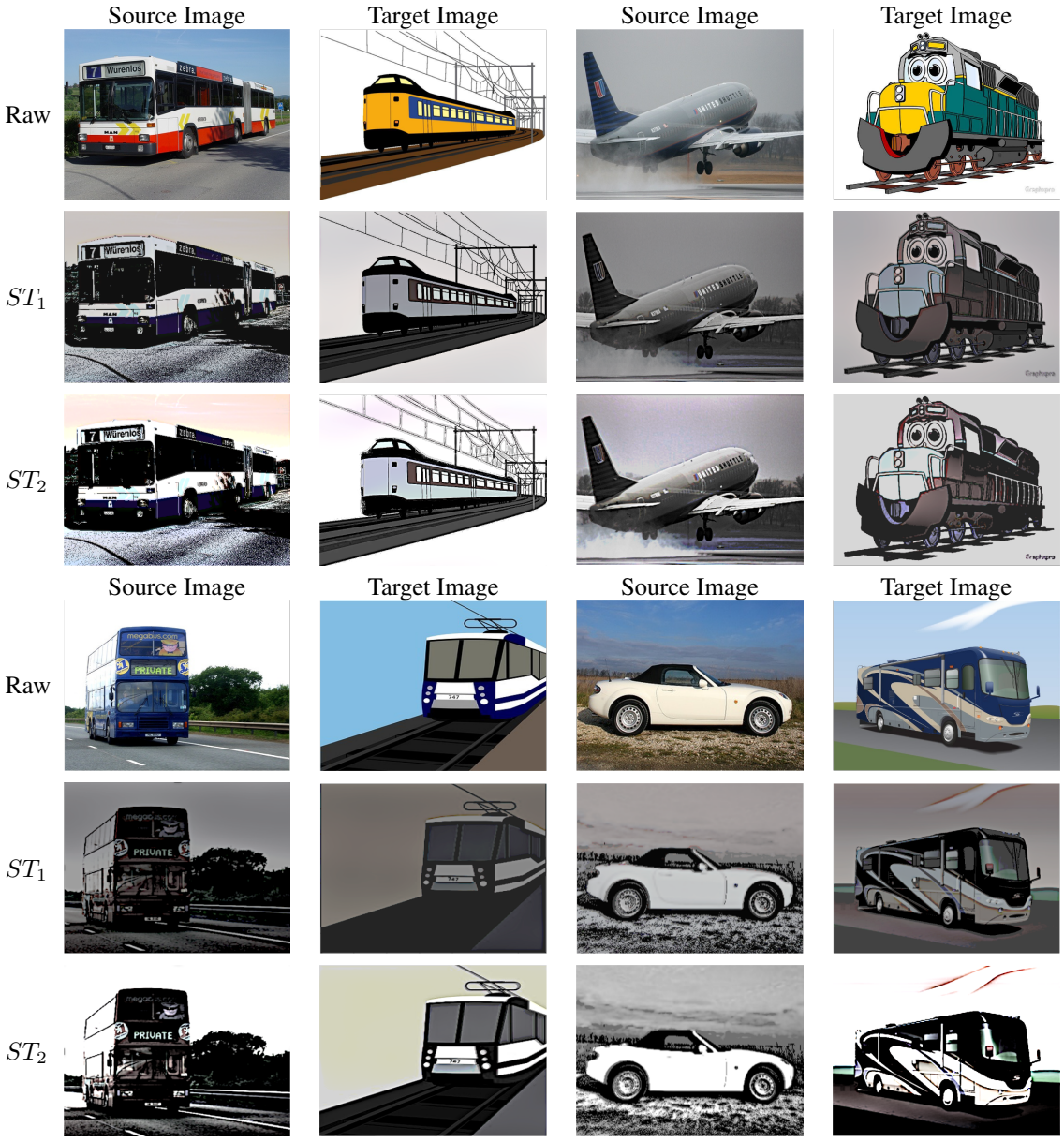

Figure 1. Visualization of ST-generated images over UDA-based object detection task PASCAL VOC → Clipart1k: For paired images of the two domains shown in rows 1 and 4 (denoted by 'Raw'), the ST_1 and ST_2 generated images (denoted by 'ST_1' and 'ST_2') in rows 2-3 and 5-6 have clearly lower inter-domain discrepancy which facilitates the inter-domain adaptation greatly.

B.2. Comparison with Existing Spectrum-based Techniques

We compared SUDA with two existing spectrum-based UDA techniques [9, 21]. RDA [9] tackles UDA problem by mitigating the overfitting problem. Specifically, it employs Fourier adversarial attacking to prevent over minimization of supervised and unsupervised UDA losses. FDA [21] tackles UDA problem by mitigating inter-domain discrepancy in an unlearnable manner, where it swaps certain pre-defined FCs of source and target samples to generate target-style source images. As a comparison, the proposed SUDA minimizes inter-domain discrepancy by identifying and enhancing domain-invariant FCs in a learnable way. In addition, SUDA introduces multi-view spectral learning for capturing more diverse target representations. SUDA thus addresses the UDA challenges from very different perspectives which is clearly complementary to the two spectrum-based works. Additionally, FDA conducted three rounds of self-training while SUDA just conducted one round,

Figure 2. Visualization of ST-generated images over UDA-based object detection task Cityscapes → Foggy cityscapes: For paired images of the two domains shown in rows 1 and 4 (denoted by 'Raw'), the ST_1 and ST_2 generated images (denoted by 'ST_1' and 'ST_2') in rows 2-3 and 5-6 have clearly lower inter-domain discrepancy which facilitates the inter-domain adaptation greatly.

	Number of ST modules			
Method	1	2	3	4
SUDA (w/o MSL)	40.6	41.8	41.6	41.7

Table 1. The number of ST affects UDA-based object detection task Cityscapes → Foggy cityscapes.

Method	1-round	3-round
FDA	46.8	50.5
SUDA	48.8	51.6

Table 2. Comparison with FDA [21] under similar training settings on UDA-based semantic segmentation task GTA5 → Cityscapes.

where multi-round self-training usually boosts performance in UDA tasks. Here we tested SUDA and FDA under similar self-training settings and the table below shows experimental results on GTA-to-Cityscapes. As table 2 shows, SUDA outperforms FDA consistently under similar self-training settings.

B.3. Comparison with Other Image Translation Method

We performed new experiments by replacing the proposed ST with CycleGAN [74] (widely used for image translation and style transfer) with the rest components unchanged. Table 3 shows experimental results on task Cityscapes → Foggy cityscapes. It can be seen that ST outperforms CycleGAN clearly as it disentangles image signals to different frequency bands and aligns them across domains separately.

B.4. Parameter Studies

Parameter N: Parameter N decides the number of the decomposed frequency components for each input image. We studied the sensitivity of N by changing it from 16 to

	Cycle-GAN	ST
mAP	37.8	42.8

Table 3. Comparison with other image translation method under UDA-based object detection task Cityscapes → Foggy cityscapes.

	N (the number of frequency components)					
Method	16	24	32	40	48	56
SUDA	41.5	41.8	42.8	42.5	42.7	42.5

Table 4. The sensitivity of parameter N affects UDA-based object detection task Cityscapes → Foggy cityscapes.

	λ_c				
λ_s	0.01	0.05	0.1	0.5	1.0
0.1	41.3	42.7	42.8	42.1	41.7

Table 5. The sensitivity of balance weights λ_c affects UDA-based object detection task Cityscapes → Foggy cityscapes.

	λ_s				
λ_c	0.01	0.05	0.1	0.5	1.0
0.1	41.6	42.5	42.8	42.6	42.3

Table 6. The sensitivity of balance weights λ_s affects UDA-based object detection task Cityscapes → Foggy cityscapes.

56 with a step of 8. Table 4 shows experimental results over UDA-based object detection task Cityscapes → Foggy cityscapes (using deformable-DETR [22]). It can be seen that the detection performance is quite tolerant to the parameter N and the best performance is obtained when $N = 32$.

Balance Weights: The weights λ_c and λ_s balance the influences of inter-domain adaptation loss $\mathcal{L}_{adv}$ and self-supervised learning loss $\mathcal{L}_{self}$. Here we study the sensitivity of λ_c and λ_s over UDA-based object detection Cityscapes → Foggy Cityscapes (using deformable-DETR [22]).

First, we fix λ_s at 0.1 and change λ_c from 0.01 to 1.0. As shown in Table 5, the detection performance is quite tolerant to λ_c and the best detection performance is obtained when λ_c is set at 0.1. In addition, we fix λ_c at 0.1 and change λ_s from 0.01 to 1.0. as shown in Table 6. It can be seen that the detection performance is quite tolerant to λ_s as well and the best detection performance is obtained when λ_s is set at 0.1.

C. Qualitative Results

C.1. Visualization of ST-generated Images

We present visual illustrations of ST-generated images ($i.e.$, ST_1 output and ST_2 output) over UDA-based object detection tasks Pascal VOC → Clipart1k and Cityscapes → Foggy cityscapes. Figs. 1 and 2 show the corresponding illustrations, respectively. It can be seen that the ST-generated source and target images have smaller inter-domain discrepancy which is desirable in UDA-based object detection tasks.

C.2. Qualitative Detection Results

We present qualitative illustrations and comparisons over UDA-based object detection task Cityscapes → Foggy cityscapes. We compare the proposed SUDA and state-of-the-art method SAP [12] over standard deformable-DETR [22]. As Fig. 3 shows, deformable-DETR produces a number of false detections due to the domain gap. SAP generates more precise bounding boxes but misses some small objects. The proposed SUDA adapts from normal weather to foggy weather well and can detect more valid objects under fogs.

References

[1] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010. 1

[2] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11457–11466, 2019. 1

[3] Lluis Castrejon, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Learning aligned cross-modal representations from weakly aligned data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2940–2949, 2016. 1

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 1

[5] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018. 1

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1

[7] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 1

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceed-*

DETR(Baseline) [22] SAP [12] **SUDA(ours)** Ground Truth

Figure 3. Qualitative illustration of domain adaptive object detection over the task Cityscapes → Foggy cityscapes: The proposed SUDA adapts well from normal weather to foggy weather and detects various challenging object such as small objects and occluded objects accurately. Deformable-DETR [22] and SAP [12] do not mitigate the domain gaps well which produce sub-optimal cross-domain alignment and object detection.

ings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. 1

[9] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Rda: Robust domain adaptation via fourier adversarial attacking. *arXiv preprint arXiv:2106.02874*, 2021. 2

[10] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5001–5009, 2018. 1

[11] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain

adaptive representation learning paradigm for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12456–12465, 2019. 1

[12] Congcong Li, Dawei Du, Libo Zhang, Longyin Wen, Tiejian Luo, Yanjun Wu, and Pengfei Zhu. Spatial attention pyramid network for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 481–497. Springer, 2020. 4, 5

[13] Xingchao Peng, Ben Usman, Neela Kaushik, Dequan Wang, Judy Hoffman, and Kate Saenko. Visda: A synthetic-to-real benchmark for visual domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2021–2026, 2018. 1

[14] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015. 1

[15] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016. 1

[16] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 1

[17] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010. 1

[18] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6956–6965, 2019. 1

[19] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, 2018. 1

[20] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7472–7481, 2018. 1

[21] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020. 2, 3

[22] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1, 4, 5

[23] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. 1

[24] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5982–5991, 2019. 1