

Spectral Unsupervised Domain Adaptation for Visual Recognition

Jingyi Zhang^{1,2,*} Jiaxing Huang^{2,*} Zichen Tian¹ Shijian Lu^{1,2,†}

¹ S-lab ² School of Computer Science and Engineering, Nanyang Technological University

{Jingyi.Zhang, Jiaxing.Huang, Zichen.Tian, Shijian.Lu}@ntu.edu.sg

Abstract

Though unsupervised domain adaptation (UDA) has achieved very impressive progress recently, it remains a great challenge due to missing target annotations and the rich discrepancy between source and target distributions. We propose Spectral UDA (SUDA), an effective and efficient UDA technique that works in the spectral space and can generalize across different visual recognition tasks. SUDA addresses the UDA challenges from two perspectives. First, it introduces a spectrum transformer (ST) that mitigates inter-domain discrepancies by enhancing domain-invariant spectra while suppressing domain-variant spectra of source and target samples simultaneously. Second, it introduces multi-view spectral learning that learns useful unsupervised representations by maximizing mutual information among multiple ST-generated spectral views of each target sample. Extensive experiments show that SUDA achieves superior accuracy consistently across different visual tasks in object detection, semantic segmentation and image classification. Additionally, SUDA also works with the transformer-based network and achieves state-of-the-art performance on object detection.

1. Introduction

Deep learning techniques [29,44,71] have achieved great success in various visual recognition tasks such as image classification [29,44,71], image segmentation [3,10,51,63] and object detection [8,20,21,50,60,61]. The great success is at the price of large quantities of annotated training data which are often prohibitively laborious and time-consuming to collect [14,15,18,49]. One alternative that could mitigate this constraint is to leverage the off-the-shelf labeled data from one or multiple related *source domains*. However, the model trained with source-domain data often experiences clear performance drop while applied to a *target domain* where the data often have discrepant distributions as compared with the source-domain data [13,68,72].

*Equal contribution

†Corresponding author

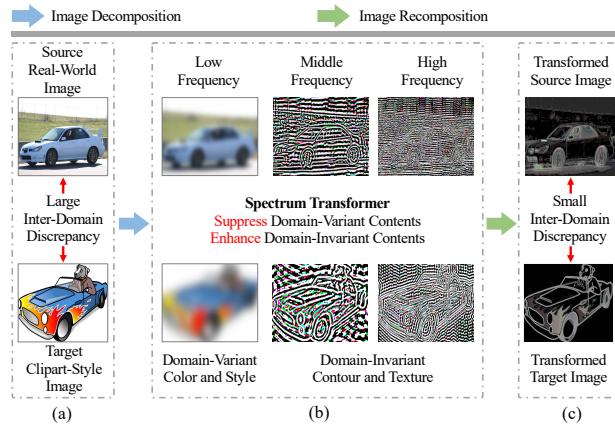


Figure 1. Illustration of the proposed spectrum transformer (ST): For images of different domains with clear distribution discrepancies as shown in (a), ST converts them into frequency space and decomposes the converted frequency signals into multiple frequency components (FCs) in low, middle, and high frequency bands as shown in (b). It learns to identify and enhance domain-invariant FCs and suppress domain-variant FCs which effectively mitigates the inter-domain discrepancy as shown in (c). Note we increase the image contrast for better visualizing (c).

Unsupervised domain adaptation (UDA) has been explored to mitigate the discrepancy between source and target domains. One typical approach is image-to-image translation with generative adversarial networks (GANs) which aligns source and target data in the input space by modifying source data to have similar styles as target data [40,43,48]. However, image-to-image translation needs to learn large amounts of parameters which is usually computationally intensive. In addition, it impairs the end-to-end feature of UDA as it needs to train GANs first before applying them for image translation [40,43,48]. Further, it could degrade UDA by undesirably modifying domain-invariant image structures that are closely entangled with domain-variant image styles in the spatial space [23,89].

We propose Spectral UDA (SUDA) that tackles UDA challenges by learning domain-invariant spectral features efficiently and effectively. SUDA works from two per-

spectives. First, it introduces a spectrum transformer (ST) that learns to reduce inter-domain discrepancies by enhancing domain-invariant frequency components (FCs) and suppressing domain-variant FCs as illustrated in Fig. 1. To this end, we design novel adversarial spectrum attention (ASA) that can identify domain-variant and domain-invariant FCs accurately. Second, we design multi-view spectral learning (MSL) that learns diverse target representations by maximizing the mutual information among multiple ST-generated spectral views for each target sample. MSL introduces certain self-supervision which mitigates the lack of target annotations effectively.

The proposed SUDA has three desirable features. First, it is generalizable and performs consistently well across different visual tasks such as image classification, image segmentation and object detection. Second, it is an online and learnable technique whereas GANs-based image translation is offline and traditional image preprocessing is mostly non-learnable. Third, it is complementary with existing UDA methods and can be incorporated with consistent and clear performance boosts but little extra computation.

The contributions of this work are threefold. *First*, we designed SUDA that tackles UDA challenges effectively by learning domain-invariant spectral features. *Second*, we design an online learnable spectrum transformer that mitigates inter-domain discrepancy by enhancing domain-invariant FCs and suppressing domain-variant FCs simultaneously. To this end, we design ASA that leverages contextual information to identify domain-variant and domain-invariant FCs accurately. *Third*, we design MSL that can learn diverse target representations by maximizing mutual information among multiple spectral views of each target sample. MSL mitigates the lack of target annotations effectively.

2. Related Works

Unsupervised Domain Adaptation. UDA has been studied extensively in recent years, largely for alleviating data annotation constraint in deep network training in various visual recognition tasks [13, 19, 26–28, 33, 35, 38, 39, 53, 75, 92]. Besides adversarial learning [13, 30, 46, 67, 72, 74, 88] and self-training [42, 84, 92, 93], image-to-image translation [40, 43, 48] has been studied for reducing inter-domain discrepancy in the input space. To this end, a number of GANs [22, 89] have been designed for translating image styles yet with minimal modification of image structures. However, GAN training is usually time-consuming, which generally makes UDA frameworks not end-to-end trainable as they need to train GANs first before applying them to UDA. In addition, GAN-based translation works in spatial space where image styles and image structures are closely entangled which inevitably modifies image structures undesirably. Some work [83] attempts to translate images in frequency space by swapping certain pre-defined

FCs of source and target images, but it is non-learnable and cannot accommodate individual images that usually have different spectral characteristics. Recently, a few studies handle UDA via contrastive learning [36, 86] and adversarial attacking [37].

We design a spectrum transformer that learns to identify domain-variant and domain-invariant FCs for each individual image. It mitigates the inter-domain discrepancy by adaptively enhancing domain-invariant FCs and suppressing domain-variant FCs across source and target images.

Learning in Frequency Space. Image preprocessing in spectral space has been widely studied with various spectral filters in the traditional image processing studies [4, 9]. However, most traditional spectral preprocessing techniques are deterministic which handle each individual image in the same manner. Spectrum learning has attracted increasing attention recently with the advance of deep learning, and it has been studied for different vision tasks such as image translation [5, 17, 80], image compression [80], network generalization [34, 81] and domain adaptation [37, 82].

We explored spectrum learning for the task of UDA. Specifically, we design multi-view spectral learning that generates different spectral views for each target image and maximizes their mutual information to learn diverse target representations without any image labels or annotations.

Visual Attention. Visual attention has been widely studied in various visual recognition tasks. It can be broadly categorized into channel attention [32, 59] and spatial attention [76] that aim to identify informative channels and spatial dependencies within each single channel, respectively. Certain hybrid attention [7, 47, 77] which combines channel and spatial attention has also been developed for better focus on informative image regions. Recently, self-attention [73] has attracted increasing interest due to its powerful capability in learning spatial dependencies within input images. In addition, self-attention, which aggregates information across multiple self-attention at different image positions, has been explored in different visual recognition tasks [8, 16, 78, 87, 91].

We design ASA that introduces adversarial learning to help identify domain-variant and domain-invariant FCs. ASA works with multiple disentangled FCs which allows to model attention effectively.

Other Related Works. Our proposed multi-view spectral learning is also related to consistency training [1, 12, 45, 55, 56], which enforces prediction consistency between different views of an input image. For example, PixMatch [55] directly enforces pixel-level consistency of predictions from different data augmentations (*e.g.*, CutMix, Fourier augmentation). SAC [1] acquires robust pseudo labels by fusing multiple predictions from different scales and retrains the network with data augmentations.

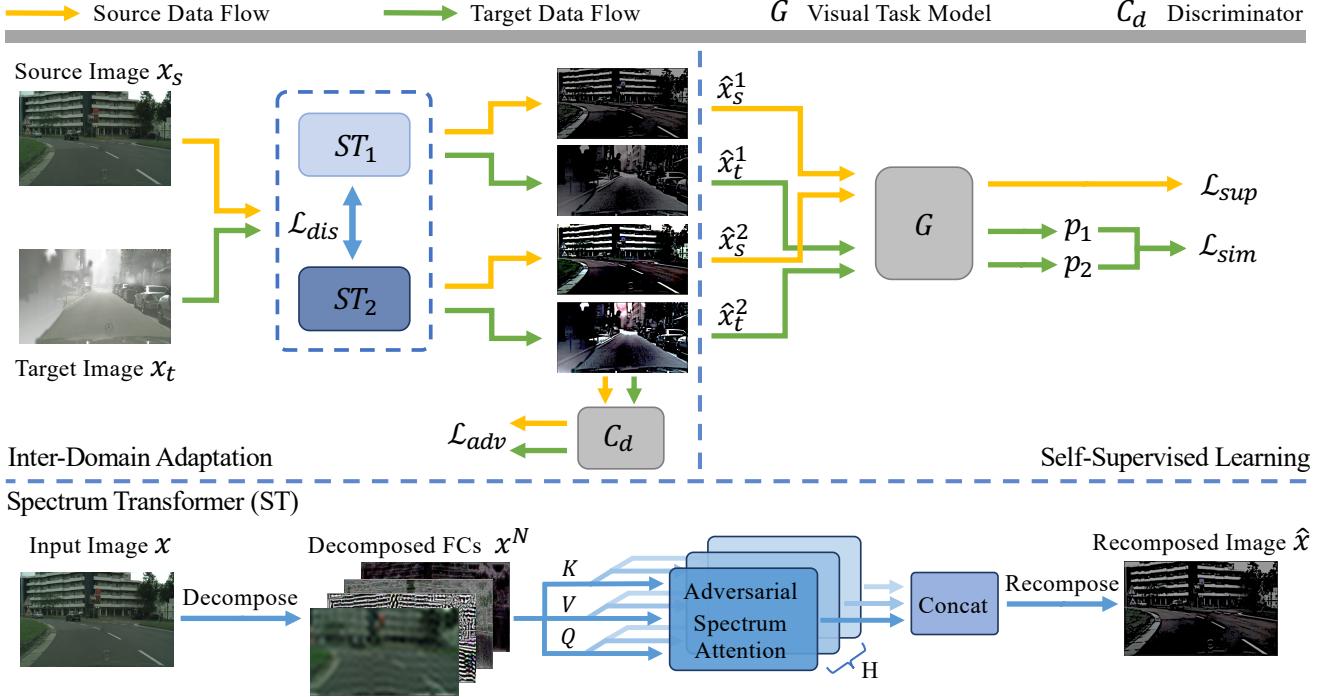


Figure 2. The overview of the proposed SUDA: For source image x_s and target image x_t , SUDA first creates two spectral views for each of them with ST_1 and ST_2 (having different parameters via a discrepancy loss \mathcal{L}_{dis}) and then feeds the spectral views to a discriminator C_d for adversarial learning for inter-domain adaptation. The ST outputs \hat{x}_s^1 and \hat{x}_s^2 are fed to a visual task model G for supervised learning, while \hat{x}_t^1 and \hat{x}_t^2 are fed to G for self-supervised learning under an unsupervised similarity loss \mathcal{L}_{sim} . The graph at the bottom shows more details of the ST design. For an input image x , ST first transforms it to spectral representations which are then decomposed to N FCs x^N . ASA then learns to identify and enhance domain-invariant FCs and suppress domain-variant FCs adaptively. Finally, the re-weighted FCs are recomposed back to a spatial-space image \hat{x} for the ensuing supervised and self-supervised learning.

3. Method

3.1. Task Definition

This work focuses on UDA in different visual recognition tasks such as image classification, image segmentation, and object detection. It involves a labeled source domain $\mathcal{D}_s = \{(x_s^i, y_s^i)\}_{i=1}^{N_s}$, where y_s^i is the label of the sample x_s^i , and an unlabeled target domain $\mathcal{D}_t = \{(x_t^i)\}_{i=1}^{N_t}$. The goal is to train a model G that well performs in \mathcal{D}_t . The baseline model is trained with the data in \mathcal{D}_s only:

$$\mathcal{L}_{sup} = l(G(x_s), y_s), \quad (1)$$

where $l(\cdot)$ denotes a task-related loss, e.g., the standard cross-entropy loss for image classification.

3.2. Spectral Unsupervised Domain Adaptation

We propose SUDA, an innovative spectral-space UDA technique that handles UDA by learning domain-invariant spectral features. SUDA has two key designs including a spectrum transformer for inter-domain adaptation and multi-view spectral learning for self-supervised learning.

Overview. Fig. 2 shows the framework of the proposed SUDA and the design of ST. Given a source-domain image $x_s \in \mathcal{D}_s$ and a target-domain image $x_t \in \mathcal{D}_t$, two complementary STs ST_1 and ST_2 (having different parameters via a discrepancy loss \mathcal{L}_{dis}) first transform the two images into spectral space and decompose them into multiple frequency components (FCs). The proposed ASA within ST then learns to identify and enhance domain-invariant FCs and suppress domain-variant FCs simultaneously via an adversarial loss \mathcal{L}_{adv} with the discriminator C_d . ASA thus mitigates inter-domain discrepancies and leads to inter-domain adaptation effectively.

Here \hat{x}_t^1 and \hat{x}_t^2 as the output of ST_1 and ST_2 capture different spectral views of the target sample x_t . They are fed to a visual task model G for self-supervised learning, where the proposed MSL strives to maximize the mutual information among the two augmentations of x_t . Note we employ two STs for producing different spectral views of x_t which allows to learn more diverse target representations. SUDA can work with a single ST without MSL as well, more details to be discussed in the experiment section.

Spectrum Transformer: We design spectrum trans-

Algorithm 1 The proposed SUDA.

Require: Source domain \mathcal{D}_s ; Target domain \mathcal{D}_t ; Visual task model G ; Spectrum transformers ST_1 and ST_2

Ensure: Learnt networks ST_1 , ST_2 and G

- 1: **for** $iter = 1$ **to** Max_Iter **do**
- 2: Sample a source data $\{x_s, y_s\} \in \mathcal{D}_s$ and a target data $x_t \in \mathcal{D}_t$
- 3: **Inter-domain Adaptation:**
- 4: Calculate $ST_1(x_s)$, $ST_2(x_s)$, $ST_1(x_t)$ and $ST_2(x_t)$ by Eq. 2
- 5: Calculate \mathcal{L}_{adv} by Eq. 5
- 6: **Self-supervised Learning:**
- 7: Calculate \mathcal{L}_{dis} by Eq. 6
- 8: Calculate \mathcal{L}_{sim} by Eq. 7
- 9: **Supervised Learning:**
- 10: Calculate \mathcal{L}_{sup} by Eq. 1
- 11: Optimize ST_1 , ST_2 and visual task model G by Eq. 8
- 12: **end for**
- 13: **return** ST_1 , ST_2 and G

former for inter-domain adaptation. Given an image $x \in \mathbb{R}^{3 \times H \times W}$, ST first transforms it into spectral representation with Fast Fourier Transform. It then decomposes the spectral representation into N FCs (*i.e.*, $x^N = \{x^n\}_{n=1}^N$, $x^n \in \mathbb{R}^{3 \times 1 \times H \times W}$) evenly by using a band pass filter. The decomposed x^N is fed to ASA within ST to enhance domain-invariant FCs and suppress domain-variant FCs adaptively. The design of ASA is shown at the bottom of Fig. 2 and its definition is presented below.

Definition 1 The proposed ASA is defined by:

$$ASA(x^N) = \mathcal{W}_{cs}(Concat(A_1(x_*^N), A_2(x_*^N), \dots, A_h(x_*^N), \dots, A_H(x_*^N)) \mathbf{P}_H) x^N \quad (2)$$

where $x_*^N \in \mathbb{R}^{3N \times 1}$ is pooled vector of x^N , $A_h(\cdot)$ is single attention head and $Concat(\cdot)$ denotes the concatenation of the outputs of $A_h(\cdot)$. $\mathbf{P}_H \in \mathbb{R}^{Hd_h \times d}$ projects the concatenation, where $d = 3N$ and $d_h = d/H$. Channel and spatial-wise attention $\mathcal{W}_{cs}(\cdot)$ takes the projected concatenation as input and further weights the decomposed x^N . Each single attention head $A_h(\cdot)$ is defined as a standard scaled dot-product attention, which maps a query (Q) and a set of key-value (K, V) pairs into an output:

$$(K, V, Q) = x_*^N \mathbf{P}_{kvq}, \quad (3)$$

$$A_h(x_*^N) = Softmax(QK^T / \sqrt{d_h})V, \quad (4)$$

where the values of K , V and Q for each head are projected from input x_*^N by $\mathbf{P}_{kvq} \in \mathbb{R}^{d \times 3d_h}$.

In final, the output of the ASA is reshaped back to the size of $3 \times N \times H \times W$ and further recomposed to a full-spectrum spatial image $\hat{x} \in \mathbb{R}^{3 \times H \times W}$ by concatenation. The output \hat{x} from ST and its corresponding domain label (0 or 1) are then forwarded to discriminator C_d for reducing inter-domain discrepancies in the input space. C_d performs adversarial learning with an adversarial loss \mathcal{L}_{adv} :

$$\begin{aligned} \mathcal{L}_{adv} = & \mathbb{E}[\log C_d(ST(x_s))] \\ & + \mathbb{E}[\log(1 - C_d(ST(x_t)))] \end{aligned} \quad (5)$$

Remark 1 Note ST performs inter-domain adaptation by employing attention mechanisms which are essentially simple matrix multiplication operations. Compared with GANs that perform image translation, ST is much more efficient as it can be trained with visual task model G in an end-to-end manner. Specifically, ST only involves 1 attention layer with about 37,000 parameters whereas a typical image translation GAN involves 9 convolutional layers with about 11,000,000 parameters [89]. Due to the high computation costs of GANs, most GAN-based UDA methods [2, 40, 43] first train image translation GANs separately which ruins the end-to-end property of UDA undesirably.

Multi-View Spectral Learning: We develop multi-view spectral learning that exploits self-supervision for learning unsupervised target representation. For each target image x_t , SUDA creates two complementary spectral views \hat{x}_t^1 and \hat{x}_t^2 by employing two spectrum transformers ST_1 and ST_2 . We enforce ST_1 and ST_2 to have different parameters by a discrepancy weight loss \mathcal{L}_{dis} so that ST_1 and ST_2 can learn complementary domain-invariant FCs of x_t :

$$\mathcal{L}_{dis} = \frac{\vec{\theta}_1 \cdot \vec{\theta}_2}{\|\vec{\theta}_1\| \|\vec{\theta}_2\|}, \quad (6)$$

where $\vec{\theta}_1$ and $\vec{\theta}_2$ denote the parameters of ST_1 and ST_2 .

The two complementary spectral views of x_t are then forwarded to a visual task model G which produces predictions $p_1 = G(\hat{x}_t^1)$ and $p_2 = G(\hat{x}_t^2)$. To maximize the mutual information [11, 25] of the two spectral views, we minimize a similarity loss \mathcal{L}_{sim} between p_1 and p_2 by:

$$\mathcal{L}_{sim} = \|p_1 - p_2\|. \quad (7)$$

Remark 2 Note we implement two complementary STs in SUDA for learning more diverse domain-invariant spectral information. We also introduce a discrepancy loss \mathcal{L}_{dis} in Eq. 6 that enhances this feature by forcing the two STs to learn different parameters. Hence, a single ST learns to minimize inter-domain discrepancy, while the two STs learn more diverse information from different spectral views for each training sample.

Cityscapes → Foggy cityscapes Object Detection											
Methods	Backbone	person	rider	car	truck	bus	train	mcycle	bicycle	mAP	
Faster R-CNN [61] (Baseline)	ResNet-50	26.9	38.2	35.6	18.3	32.4	9.6	25.8	28.6	26.9	
DAF [13]	ResNet-50	29.2	40.4	43.4	19.7	38.3	28.5	23.7	32.7	32.0	
+SUDA	ResNet-50	39.5	46.8	54.6	29.3	50.7	44.6	31.6	39.5	42.1	
SCDA [90]	ResNet-50	33.8	42.1	52.1	26.8	42.5	26.5	29.2	34.5	35.9	
+SUDA	ResNet-50	39.7	47.7	54.3	27.6	51.8	46.5	31.2	39.6	42.3	
SWDA [67]	ResNet-50	31.8	44.3	48.9	21.0	43.8	28.0	28.9	35.8	35.3	
+SUDA	ResNet-50	39.5	48.2	57.8	29.5	52.9	37.5	34.5	41.3	42.7	
SUDA	ResNet-50	40.2	47.9	54.6	28.5	49.5	39.2	33.8	41.5	41.9	
DETR [91] (Baseline)	ResNet-50	43.7	38.0	57.2	15.2	34.7	14.4	26.1	42.4	34.0	
DAF [13]	ResNet-50	49.4	49.7	62.1	23.6	43.8	21.6	31.3	43.1	40.6	
+SUDA	ResNet-50	50.5	51.7	64.1	26.7	48.5	14.2	38.1	49.5	42.9	
SWDA [67]	ResNet-50	49.0	49.0	61.4	23.9	43.1	22.9	31.0	45.2	40.7	
+SUDA	ResNet-50	50.7	50.3	67.3	22.3	45.2	27.4	34.0	48.9	43.3	
CRDA [79]	ResNet-50	49.8	48.4	61.9	22.3	40.7	30.0	29.9	45.4	41.1	
+SUDA	ResNet-50	52.3	51.6	66.7	30.4	47.1	11.9	36.8	48.7	43.2	
CF [88]	ResNet-50	49.6	49.7	62.6	23.3	43.4	27.4	30.2	44.8	41.4	
+SUDA	ResNet-50	51.2	51.4	68.5	25.3	48.0	26.5	33.8	49.9	44.3	
SAP [46]	ResNet-50	49.3	49.9	62.5	23.0	44.1	29.4	31.3	45.8	41.9	
+SUDA	ResNet-50	51.4	52.2	67.5	28.7	49.6	28.7	39.2	50.4	46.0	
SUDA	ResNet-50	50.5	51.7	64.1	26.7	48.5	13.1	38.1	49.5	42.8	

Table 1. Experiments on UDA-based object detection task Cityscapes → Foggy Cityscapes.

PASCAL VOC → Clipart1k Object Detection																					
Methods	aero	bicyc.	bird	boat	bott.	bus	car	cat	chair	cow	table	dog	horse	bike	pers.	plant	sheep	sofa	train	tv	mAP
Baseline [61]	35.6	52.5	24.3	23.0	20.0	43.9	32.8	10.7	30.6	11.7	13.8	6.0	36.8	45.9	48.7	41.9	16.5	7.3	22.9	32.0	27.8
DAF [13]	15.0	34.6	12.4	11.9	19.8	21.1	23.2	3.1	22.1	26.3	10.6	10.0	19.6	39.4	34.6	29.3	1.0	17.1	19.7	24.8	19.8
+SUDA	28.2	53.8	37.1	15.4	37.6	66.6	35.3	21.7	38.7	48.7	18.3	28.4	24.4	82.4	61.0	44.5	11.9	34.4	49.5	59.7	39.9
SWDA [67]	26.2	48.5	32.6	33.7	38.5	54.3	37.1	18.6	34.8	58.3	17.0	12.5	33.8	65.5	61.6	52.0	9.3	24.9	54.1	49.1	38.1
+SUDA	33.7	61.8	36.9	23.1	39.2	56.2	33.9	23.4	38.7	45.9	15.4	23.4	25.8	75.8	58.6	41.8	15.7	33.2	61.7	60.1	40.2
SUDA	33.8	56.1	32.3	24.1	30.9	54.6	38.5	18.4	34.0	41.0	18.7	24.3	29.5	84.0	57.6	49.1	14.3	36.3	55.6	51.6	39.2

Table 2. Experiments on UDA-based object detection task PASCAL VOC → Clipart1k.

Cityscapes → Foggy cityscapes					
Methods	L_{sup}	L_{adv}	L_{dis}	L_{sim}	mAP
Baseline [91]	✓				34.0
+Single ST	✓	✓			40.6
+Two STs	✓	✓	✓		41.8
+Two STs +MSL(SUDA)	✓	✓	✓	✓	42.8

Table 3. Ablation study of the proposed Spectrum Transformer and Multi-view Spectral Learning over object detection task Cityscapes → Foggy Cityscapes.

Overall Training Objective. The objective of SUDA consists of three losses as stated in Algorithm 1, namely, the supervised task loss \mathcal{L}_{sup} in Eq. 1, the inter-domain adaptation loss \mathcal{L}_{adv} in Eq. 5, and the self-supervised learning loss \mathcal{L}_{self} which consists of the discrepancy loss \mathcal{L}_{dis} and the similarity loss \mathcal{L}_{sim} in Eqs. 6 and 7. The overall training objective can thus be formulated by

$$\max_{C_d} \min_{G, ST} \mathcal{L}_{sup} - \lambda_c \mathcal{L}_{adv} + \lambda_s \mathcal{L}_{self}, \quad (8)$$

where λ_c and λ_s denote the balance weights.

4. Experiment

This section presents experiments including datasets and implementation details, domain adaptation evaluations for object detection, image classification, and semantic segmentation tasks, and discussion, respectively. More details are to be described in the ensuing subsections.

4.1. Datasets

We evaluate SUDA over multiple datasets across different visual UDA tasks on object detection, image classification and semantic segmentation as listed:

UDA for Object Detection: We study two object detection tasks Cityscapes [14] → Foggy Cityscapes [69] and PASCAL VOC [18] → Clipart1k [40].

UDA for Image Classification: We study two UDA-based image classification tasks VisDA17 [57] and Office-31 [65]. For VisDA17, we evaluate the task synthetic→real. For Office-31, we study six adaptation tasks: A→W, D→W, W→D, A→D, D→A, and W→A.

UDA for Semantic Segmentation: We study two synthetic-to-real semantic segmentation tasks GTA5 [62] →

VisDA17 Classification													
Methods	aero.	bike	bus	car	horse	knife	motor	person	plant	skate.	train	truck	Mean
Res-101 [29]	55.1	53.3	61.9	59.1	80.6	17.9	79.7	31.2	81.0	26.5	73.5	8.5	52.4
MCD [68]	87.0	60.9	83.7	64.0	88.9	79.6	84.7	76.9	88.6	40.3	83.0	25.8	71.9
ADR [66]	87.8	79.5	83.7	65.3	92.3	61.8	88.9	73.2	87.8	60.0	85.5	32.3	74.8
SimNet-Res152 [58]	94.3	82.3	73.5	47.2	87.9	49.2	75.1	79.7	85.3	68.5	81.1	50.3	72.9
GTA-Res152 [70]	-	-	-	-	-	-	-	-	-	-	-	-	77.1
CBST [92]	87.2	78.8	56.5	55.4	85.1	79.2	83.8	77.7	82.8	88.8	69.0	72.0	76.4
+SUDA	89.6	79.0	69.0	66.1	88.5	79.9	86.7	79.6	85.4	87.7	81.0	73.8	80.5
CRST [93]	88.0	79.2	61.0	60.0	87.5	81.4	86.3	78.8	85.6	86.6	73.9	68.8	78.1
+SUDA	91.5	79.7	71.9	66.5	88.5	81.1	85.6	79.5	86.2	86.5	79.9	74.3	80.9
SUDA	88.3	79.3	66.2	64.7	87.4	80.1	85.9	78.3	86.3	87.5	78.8	74.5	79.8

Table 4. Experiments on UDA-based image classification task VisDA17.

Office-31 Classification												
Methods	A→W	D→W	W→D	A→D	D→A	W→A	Mean					
ResNet-50 [29]	68.4	96.7	99.3	68.9	62.5	60.7	76.1					
JAN [52]	85.4	97.4	99.8	84.7	68.6	70.0	84.3					
GTA [70]	89.5	97.9	99.8	87.7	72.8	71.4	86.5					
CBST [92]	87.8	98.5	100.0	86.5	71.2	70.9	85.8					
+SUDA	90.5	98.6	100.0	91.4	72.7	72.1	87.6					
CRST [93]	89.4	98.9	100.0	88.7	72.6	70.9	86.8					
+SUDA	91.0	98.8	100.0	91.9	72.9	72.3	87.8					
SUDA	90.8	98.7	100.0	91.2	72.2	71.4	87.4					

Table 5. Experiments on UDA-based image classification task Office-31.

Cityscapes [14] and SYNTHIA [64] → Cityscapes.

Due to the space limit, we provide more details about datasets in Section A.1 in supplementary materials.

4.2. Implementation Details

Object Detection: For Cityscapes→ Foggy Cityscapes, we adopt Faster R-CNN [61] and deformable-DETR [91] as detection networks and ResNet-50 [29] as backbone as in [6,91]. For PASCAL VOC → Clipart1k, we adopt Faster R-CNN with ResNet-101 [29] as in [40,67].

Image Classification: Following [65, 93], we use ResNet-101 and ResNet-50 [29] as backbones for the tasks VisDA17 and Office-31, respectively.

Semantic Segmentation: We use DeepLab-V2 [10] with ResNet-101 [29] as the segmentation network as in [72,92].

For all visual recognition tasks, we set the number of FCs N at 32. Due to the space limit, we provide more implementation details in Section A.2 in supplementary materials.

4.3. Domain Adaptive Object Detection

We first benchmark the proposed SUDA with state-of-the-art domain adaptive object detection methods over two UDA tasks Cityscapes→ Foggy Cityscapes and PASCAL VOC → Clipart1k. Tables 1 and 2 show experimental results. It can be seen that SUDA achieves competitive object detection performance as compared with all highly-

optimized state-of-the-art methods across two very different network architectures (Faster R-CNN and deformable-DETR). In addition, SUDA is complementary to most existing methods which produces clear and consistent performance boosts while incorporated as a plug-in.

We also examine the proposed SUDA by performing several ablation studies over a domain adaptive object detection task Cityscapes→ Foggy Cityscapes. Table 3 shows experimental results. It can be seen that including either one or two STs in **+Single ST** and **+Two STs** outperforms the **Baseline** (deformable-DETR) by large margins. In addition, including two complementary STs in **+Two STs** perform clearly better than including a single ST in **+Single ST** as the two STs learn more diverse and complementary domain-invariant spectral information. Further including MSL beyond the two STs in **+Two STs +MSL (SUDA)** performs clearly the best, demonstrating the effectiveness of the proposed multi-view spectral learning.

4.4. Domain Adaptive Image Classification

We evaluate and benchmark SUDA over two domain adaptive image classification tasks VisDA17 and Office-31. Tables 4 and 5 shows experimental results, where SUDA outperforms all state-of-the-art methods clearly. In addition, SUDA is complementary to existing methods which produces consistent and clear performance boosts while incorporated as a plug-in. Note we perform the complementary studies over a few representative domain adaptive image classification methods only due to space limit.

4.5. Domain Adaptive Semantic Segmentation

We evaluate and benchmark SUDA over two domain adaptive semantic segmentation tasks GTA5 → Cityscapes and SYNTHIA → Cityscapes. Tables 6 and 7 show experimental results. We can see that SUDA achieves competitive segmentation performance as compared with highly-optimized state-of-the-art methods. In addition, it is complementary with existing methods which produce consistent performance boosts while incorporated as a plug-in.

GTA5 → Cityscapes Semantic Segmentation																				
Methods	road	side.	buil.	wall	fence	pole	light	sign	vege.	ter.	sky	pers.	rider	car	truck	bus	train	mot.	bike	mIoU
Baseline [29]	75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.5	25.3	36.0	36.6
AdaptSeg [72]	86.5	36.0	79.9	23.4	23.3	23.9	35.2	14.8	83.4	33.3	75.6	58.5	27.6	73.7	32.5	35.4	3.9	30.1	28.1	42.4
CBST [92]	91.8	53.5	80.5	32.7	21.0	34.0	28.9	20.4	83.9	34.2	80.9	53.1	24.0	82.7	30.3	35.9	16.0	25.9	42.8	45.9
AdvEnt [75]	89.4	33.1	81.0	26.6	26.8	27.2	33.5	24.7	83.9	36.7	78.8	58.7	30.5	84.8	38.5	44.5	1.7	31.6	32.4	45.5
CRST [93]	91.0	55.4	80.0	33.7	21.4	37.3	32.9	24.5	85.0	34.1	80.8	57.7	24.6	84.1	27.8	30.1	26.9	26.0	42.3	47.1
BDL [48]	91.0	44.7	84.2	34.6	27.6	30.2	36.0	36.0	85.0	43.6	83.0	58.6	31.6	83.3	35.3	49.7	3.3	28.8	35.6	48.5
CrCDA [39]	92.4	55.3	82.3	31.2	29.1	32.5	33.2	35.6	83.5	34.8	84.2	58.9	32.2	84.7	40.6	46.1	2.1	31.1	32.7	48.6
RDA [37]	89.8	39.1	81.7	27.6	19.9	34.2	35.9	23.3	82.1	29.5	76.6	58.3	26.0	82.1	32.5	45.2	15.3	26.9	33.5	45.2
+SUDA	91.5	52.1	82.2	32.3	24.2	36.2	44.3	36.3	84.1	39.4	78.3	59.6	26.2	83.7	37.5	45.8	12.4	27.7	39.0	49.1
TIR [41]	92.9	55.0	85.3	34.2	31.1	34.9	40.7	34.0	85.2	40.1	87.1	61.0	31.1	82.5	32.3	42.9	0.3	36.4	46.1	50.2
+SUDA	92.6	54.9	85.9	31	30.6	37.6	43.6	41.3	84.5	39.3	87	60.4	32.6	84.6	38.3	46.7	11.2	34.9	43.7	51.6
FDA [83]	92.5	53.3	82.4	26.5	27.6	36.4	40.6	38.9	82.3	39.8	78.0	62.6	34.4	84.9	34.1	53.1	16.9	27.7	46.4	50.5
+SUDA	93.4	55.1	84.9	31.5	28.9	38.3	45.6	41.9	84.6	40.0	83.1	61.4	31.3	84.8	41.1	50.5	15.4	30.8	43.9	51.9
ProDA [85]	87.8	56.0	79.7	46.3	44.8	45.6	53.5	53.5	88.6	45.2	82.1	70.7	39.2	88.8	45.5	59.4	1.0	48.9	56.4	57.5
+SUDA	94.5	67.5	86.4	45.1	41.4	47.1	50.5	55.6	89.6	48.1	87.4	67.3	1.1	88.9	39.1	60.2	33.3	44.5	61.1	58.3
SUDA	91.1	52.3	82.9	30.1	25.7	38.0	44.9	38.2	83.9	39.1	79.2	58.4	26.4	84.5	37.7	45.6	10.1	23.1	36.0	48.8

Table 6. Experiments on UDA-based semantic segmentation task GTA5 → Cityscapes.

SYNTHIA → Cityscapes Semantic Segmentation																			
Methods	road	side.	buil.	wall	fence	pole	light	sign	vege.	sky	pers.	rider	car	bus	mot.	bike	mIoU	mIoU*	
Baseline [29]	55.6	23.8	74.6	9.2	0.2	24.4	6.1	12.1	74.8	79.0	55.3	19.1	39.6	23.3	13.7	25.0	33.5	38.6	
AdaptSeg [72]	84.3	42.7	77.5	-	-	-	4.7	7.0	77.9	82.5	54.3	21.0	72.3	32.2	18.9	32.3	-	46.7	
AdvEnt [75]	85.6	42.2	79.7	8.7	0.4	25.9	5.4	8.1	80.4	84.1	57.9	23.8	73.3	36.4	14.2	33.0	41.2	48.0	
CrCDA [39]	86.2	44.9	79.5	8.3	0.7	27.8	9.4	11.8	78.6	86.5	57.2	26.1	76.8	39.9	21.5	32.1	42.9	50.0	
CRST [93]	67.7	32.2	73.9	10.7	1.6	37.4	22.2	31.2	80.8	80.5	60.8	29.1	82.8	25.0	19.4	45.3	43.8	50.1	
TIR [41]	92.6	53.2	79.2	-	-	-	1.6	7.5	78.6	84.4	52.6	20.0	82.1	34.8	14.6	39.4	-	49.3	
+SUDA	83.9	40.1	76.9	4.5	0.1	26.1	22.9	26.4	79.6	80.7	58.1	28.3	81.0	37.4	35.1	46.8	45.5	53.6	
FDA [83]	79.3	35.0	73.2	-	-	-	19.9	24.0	61.7	82.6	61.4	31.1	83.9	40.8	38.4	51.1	-	52.5	
+SUDA	85.6	38.8	76.7	9.2	0.2	28.4	25.4	27.0	78.4	81.7	60.4	28.6	82.8	38.8	36.2	48.1	46.7	54.5	
SUDA	83.4	36.0	71.3	8.7	0.1	26.0	18.2	26.7	72.4	80.2	58.4	30.8	80.6	38.7	36.1	46.1	44.6	52.2	

Table 7. Experiments on UDA-based semantic segmentation task SYNTHIA → Cityscapes. mIoU is evaluated on 16 classes, and mIoU* is evaluated on 13 classes.

5. Discussion

Generalization across Visual Tasks: The proposed SUDA is generally applicable to various visual recognition tasks as described in Sections 4.3, 4.4 and 4.5. With simple implementation and minimal fine-tuning as described in Section 4.2, it produces competitive performance consistently across different tasks as shown in Tables 1-7. The superior generalization is largely attributed to the spectral transformer and multi-view spectral learning which are task-agnostic by learning domain-invariant spectra.

Complementarity Study: The proposed SUDA is complementary to most existing visual recognition methods consistently (while incorporated as a plug-in) as shown in Tables 1-7. The synergistic effect is largely attributed to the proposed spectrum transformer and multi-view spectral learning which work in spectral domain whereas most existing methods work in spatial domain.

ST Analysis: We examine how the proposed ST learns to produce domain-invariant spectral information over domain adaptive object detection task Cityscapes → Foggy

cityscapes. We first study the visual features of source and target samples that are produced by ST_1 and ST_2 in Fig. 2. As Fig. 3 shows, the ST-generated source and target features are better aligned (with smaller inter-domain distances) as compared with that of the original images, showing that ST helps learn more domain-invariant features effectively. Due to space limit, we provide the visualization of ST-generated images in Section C.1 of the supplementary materials.

In addition, we analyze ST quantitatively by measuring the inter-domain distance [31] before and after the ST transformation of the source and target images. As Table 8 shows, the cross-domain distance is reduced significantly by implementing either ST_1 (under (X_s^1, X_t^1)) or ST_2 (under (X_s^2, X_t^2)) alone as compared with the cross-domain distance of the original source and target samples. While ST_1 and ST_2 are both implemented, the cross-domain distance is further reduced as shown in the column $(X_s^{1,2}, X_t^{1,2})$.

Analysis of Discrepancy Loss \mathcal{L}_{dis} : We also study how \mathcal{L}_{dis} guides ST_1 and ST_2 to learn more diverse features for both source and target samples. This study is based on

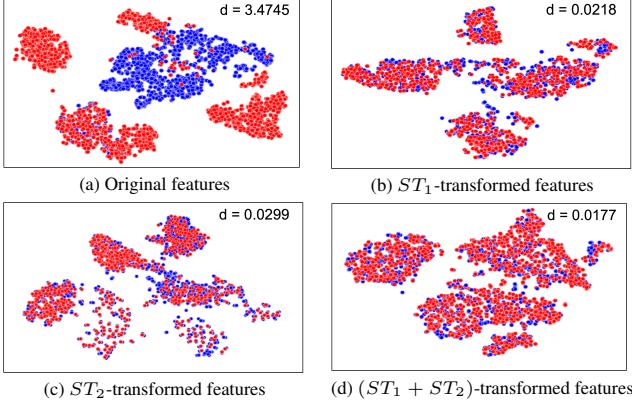


Figure 3. Visualization of feature representations via t-SNE [54]: Red points represent source features and blue points represent target features. d denotes the distance between source and target feature representations as measured by Maximum Mean Discrepancy [24]. A single ST helps to reduce inter-domain distance significantly as shown in (b) and (c). Two complementary STs can further reduce the inter-domain distance clearly as shown in (d).

intra-domain distance [31] over domain adaptive object detection task Cityscapes → Foggy cityscapes, *i.e.*, the larger the distance, the more diverse features learnt.

Specifically, we study the intra-domain distances of ST-generated source and target images while \mathcal{L}_{dis} is present (as in (X_s^1, X_s^2) and (X_t^1, X_t^2)) and absent (as in $(X_s^{1'}, X_s^{2'})$ and $(X_t^{1'}, X_t^{2'})$). As Table 9 shows, the intra-domain distances are clearly larger when \mathcal{L}_{dis} is present, demonstrating that \mathcal{L}_{dis} effectively guides the two STs to learn more diverse representations for both source and target samples.

Number of STs: The number of STs in SUDA does affect the domain adaptation performance. Our study shows that one ST can improve the domain adaptation significantly and two complementary STs can further introduce clear improvements. However, the domain adaptation saturates with more STs which instead complicates the network structures and introduces extra parameters. Due to the space limit, we provide detailed experimental results and analysis in Section B.1 in supplementary materials.

Comparisons with Existing Spectrum-based Techniques: We compared SUDA with two existing spectrum-based UDA techniques [37, 83], where [83] swaps certain pre-defined FCs of source and target samples to mitigate inter-domain discrepancy whereas [37] employs adversarial attacking to mitigate the overfitting in UDA. As a comparison, the proposed SUDA minimizes inter-domain discrepancy by identifying and enhancing domain-invariant FCs in a learnable manner. In addition, it 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

	(X_s, X_t)	(X_s^1, X_t^1)	(X_s^2, X_t^2)	$(X_s^{1,2}, X_t^{1,2})$
CDID	58.57	18.78	20.56	16.23

Table 8. Quantitative analysis of ST on CDID (cross-domain image distance): The CDID (measured in FID [31]) is greatly reduced after transformation by either ST_1 in (X_s^1, X_t^1) or ST_2 in (X_s^2, X_t^2) . The transformation by both ST_1 and ST_2 further reduces CDID clearly as in $(X_s^{1,2}, X_t^{1,2})$.

	(X_s^1, X_s^2)	(X_t^1, X_t^2)	$(X_s^{1'}, X_s^{2'})$	$(X_t^{1'}, X_t^{2'})$
IDID	30.59	29.82	7.13	9.42

Table 9. Quantitative analysis of discrepancy loss \mathcal{L}_{dis} on IDID (intra-domain image distance): With \mathcal{L}_{dis} , the IDID (measured in FID [31]) in (X_s^1, X_s^2) and (X_t^1, X_t^2) are clearly larger as compared with $(X_s^{1'}, X_s^{2'})$ and $(X_t^{1'}, X_t^{2'})$ without using \mathcal{L}_{dis} .

to the two spectrum-based works as shown in Table 6. We provide detailed comparison and analysis at Section B.2 of the supplementary materials.

Parameter Analysis: We studied the sensitivity of the number of FCs N and balance weights λ_c and λ_s defined in Eq. 8. Due to the space limit, we provide the detailed analysis in Section B.4 in supplementary materials.

6. Conclusion

This paper presents SUDA, a spectral UDA technique that addresses UDA challenges by learning domain-invariant spectral features. SUDA consists of two key designs. The first is a spectrum transformer that mitigates inter-domain discrepancy by highlighting domain-invariant spectra and suppressing domain-variant spectra in the input space. The second is multi-view spectral learning that leverage multi-view consistency for learning diverse representations for each target sample. SUDA has three unique features: 1) it is generic to various visual recognition tasks with consistently superior performance; 2) it is learnable and end-to-end trainable in various downstream tasks; 3) it complements with existing UDA methods with consistent performance boosts. Moving forwards, we will continue to investigate frequency-space learning and its applications in various downstream computer vision tasks.

Acknowledgement. This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s). This research is also partially supported by Singtel Cognitive and Artificial Intelligence Lab for Enterprises (SCALE@NTU), which is a collaboration between Singapore Telecommunications Limited (Singtel) and Nanyang Technological University (NTU) that is supported by A*STAR under its Industry Alignment Fund (LOA Award number: I1701E0013).

References

- [1] Nikita Araslanov and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15384–15394, 2021. [2](#)
- [2] Vinicius F Arruda, Thiago M Paixão, Rodrigo F Berriel, Alberto F De Souza, Claudine Badue, Nicu Sebe, and Thiago Oliveira-Santos. Cross-domain car detection using unsupervised image-to-image translation: From day to night. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019. [4](#)
- [3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. [1](#)
- [4] Sing T Bow. *Pattern recognition and image preprocessing*. CRC press, 2002. [2](#)
- [5] Mu Cai, Hong Zhang, Huijuan Huang, Qichuan Geng, Yixuan Li, and Gao Huang. Frequency domain image translation: More photo-realistic, better identity-preserving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13930–13940, 2021. [2](#)
- [6] 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. [6](#)
- [7] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. [2](#)
- [8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. [1](#), [2](#)
- [9] Jyotismita Chaki and Nilanjan Dey. *A Beginner’s Guide to Image Preprocessing Techniques*. CRC Press, 2018. [2](#)
- [10] 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](#), [6](#)
- [11] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. [4](#)
- [12] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2613–2622, 2021. [2](#)
- [13] 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](#), [2](#), [5](#)
- [14] 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](#), [5](#), [6](#)
- [15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. [1](#)
- [16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. [2](#)
- [17] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7890–7899, 2020. [2](#)
- [18] 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](#), [5](#)
- [19] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. [2](#)
- [20] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. [1](#)
- [21] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. [1](#)
- [22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. [2](#)
- [23] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014. [1](#)
- [24] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012. [8](#)
- [25] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. [4](#)

- [26] Dayan Guan, Jiaxing Huang, Shijian Lu, and Aoran Xiao. Scale variance minimization for unsupervised domain adaptation in image segmentation. *Pattern Recognition*, 112:107764, 2021. 2
- [27] Dayan Guan, Jiaxing Huang, Aoran Xiao, and Shijian Lu. Domain adaptive video segmentation via temporal consistency regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8053–8064, 2021. 2
- [28] Dayan Guan, Jiaxing Huang, Aoran Xiao, Shijian Lu, and Yanpeng Cao. Uncertainty-aware unsupervised domain adaptation in object detection. *IEEE Transactions on Multimedia*, 2021. 2
- [29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 6, 7
- [30] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6668–6677, 2019. 2
- [31] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7, 8
- [32] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 2
- [33] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Cross-view regularization for domain adaptive panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10133–10144, 2021. 2
- [34] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. *arXiv preprint arXiv:2103.02370*, 2021. 2
- [35] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. *Advances in Neural Information Processing Systems*, 34, 2021. 2
- [36] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. *arXiv preprint arXiv:2110.03374*, 2021. 2
- [37] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Rda: Robust domain adaptation via fourier adversarial attacking. *arXiv preprint arXiv:2106.02874*, 2021. 2, 7, 8
- [38] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Multi-level adversarial network for domain adaptive semantic segmentation. *Pattern Recognition*, 123:108384, 2022. 2
- [39] Jiaxing Huang, Shijian Lu, Dayan Guan, and Xiaobing Zhang. Contextual-relation consistent domain adaptation for semantic segmentation. In *European Conference on Computer Vision*, pages 705–722. Springer, 2020. 2, 7
- [40] 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, 2, 4, 5, 6
- [41] Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. *arXiv preprint arXiv:2003.00867*, 2020. 7
- [42] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6092–6101, 2019. 2
- [43] 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, 2, 4
- [44] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 1
- [45] Xin Lai, Zhuotao Tian, Li Jiang, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Semi-supervised semantic segmentation with directional context-aware consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1205–1214, 2021. 2
- [46] 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. 2, 5
- [47] Xiang Li, Xiaolin Hu, and Jian Yang. Spatial group-wise enhance: Improving semantic feature learning in convolutional networks. *arXiv preprint arXiv:1905.09646*, 2019. 2
- [48] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6936–6945, 2019. 1, 2, 7
- [49] Tsung-Yi Lin, Michael Maire, Serge Belongie, James HaysPietro, PeronaDeva Ramanan, Piotr DollárC, and Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755, 2014. 1
- [50] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 1
- [51] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1
- [52] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017. 6

- [53] Zhipeng Luo, Zhongang Cai, Changqing Zhou, Gongjie Zhang, Haiyu Zhao, Shuai Yi, Shijian Lu, Hongsheng Li, Shanghang Zhang, and Ziwei Liu. Unsupervised domain adaptive 3d detection with multi-level consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8866–8875, 2021. 2
- [54] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 8
- [55] Luke Melas-Kyriazi and Arjun K Manrai. Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12435–12445, 2021. 2
- [56] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020. 2
- [57] 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. 5
- [58] Pedro O Pinheiro. Unsupervised domain adaptation with similarity learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8004–8013, 2018. 6
- [59] Wang Qilong, Wu Banggu, Zhu Pengfei, Li Peihua, Zuo Wangmeng, and Hu Qinghua. Eca-net: Efficient channel attention for deep convolutional neural networks. 2020. 2
- [60] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1
- [61] 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, 5, 6
- [62] 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. 5
- [63] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1
- [64] 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. 6
- [65] 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. 5, 6
- [66] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Adversarial dropout regularization. *International Conference on Learning Representations*, 2017. 6
- [67] 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. 2, 5, 6
- [68] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018. 1, 6
- [69] 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. 5
- [70] Swami Sankaranarayanan, Yogesh Balaji, Carlos D Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8503–8512, 2018. 6
- [71] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1
- [72] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Ki-hyun 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, 2, 6, 7
- [73] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 2
- [74] Vibashan VS, Vikram Gupta, Poojan Oza, Vishwanath A Sindagi, and Vishal M Patel. Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4516–4526, 2021. 2
- [75] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019. 2, 7
- [76] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 2
- [77] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 2
- [78] Enze Xie, Wenhui Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and ef-

- ficient design for semantic segmentation with transformers. *arXiv preprint arXiv:2105.15203*, 2021. 2
- [79] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11724–11733, 2020. 5
- [80] Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. Learning in the frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1740–1749, 2020. 2
- [81] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14383–14392, 2021. 2
- [82] Yanchao Yang, Dong Lao, Ganesh Sundaramoorthi, and Stefano Soatto. Phase consistent ecological domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9011–9020, 2020. 2
- [83] 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, 7, 8
- [84] Fuxun Yu, Di Wang, Yinpeng Chen, Nikolaos Karanakis, Pei Yu, Dimitrios Lymberopoulos, and Xiang Chen. Unsupervised domain adaptation for object detection via cross-domain semi-supervised learning. *arXiv preprint arXiv:1911.07158*, 2019. 2
- [85] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12414–12424, 2021. 7
- [86] Yixin Zhang, Zilei Wang, and Yushi Mao. Rpn prototype alignment for domain adaptive object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12425–12434, 2021. 2
- [87] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. *arXiv preprint arXiv:2012.15840*, 2020. 2
- [88] Yangtao Zheng, Di Huang, Songtao Liu, and Yunhong Wang. Cross-domain object detection through coarse-to-fine feature adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13766–13775, 2020. 2, 5
- [89] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1, 2, 4
- [90] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 687–696, 2019. 5
- [91] 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. 2, 5, 6
- [92] 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. 2, 6, 7
- [93] 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. 2, 6, 7