Self-Adversarial Disentangling for Specific Domain Adaptation

Qianyu Zhou*, Qiqi Gu*, Jiangmiao Pang, Zhengyang Feng,
Guangliang Cheng, Xuequan Lu, Jianping Shi, Lizhuang Ma

***Abstract*—Domain adaptation aims to bridge the domain shifts between the source and target domains. These shifts may span different dimensions such as fog, rainfall, etc. However, recent methods typically do not consider explicit prior knowledge on a specific dimension, thus leading to less desired adaptation performance. In this paper, we study a practical setting called Specific Domain Adaptation (SDA) that aligns the source and target domains in a demanded-specific dimension. Within this setting, we observe the intra-domain gap induced by different domainness (i.e., numerical magnitudes of this dimension) is crucial when adapting to a specific domain. To address the problem, we propose a novel Self-Adversarial Disentangling (SAD) framework. In particular, given a specific dimension, we first enrich the source domain by introducing a domainness creator with providing additional supervisory signals. Guided by the created domainness, we design a self-adversarial regularizer and two loss functions to jointly disentangle the latent representations into domainness-specific and domainness-invariant features, thus mitigating the intra-domain gap. Our method can be easily taken as a plug-and-play framework and does not introduce any extra costs in the inference time. We achieve consistent improvements over state-of-the-art methods in both object detection and semantic segmentation tasks.**

***Index Terms*—Domain Adaptation, Semantic Segmentation, Object Detection, Autonomous Driving.**

I. Introduction

OVER the past several years, deep neural networks have brought impressive advances in many computer vision tasks, such as object detection [1]–[3] and semantic segmentation [4]–[8]. However, the model trained in a source domain will suffer from serious performance degradation when applying to a novel domain, which limits its generalization ability in complicated real-world scenarios. Annotating a large-scale dataset for each new domain is cost-expensive and time-consuming. Unsupervised domain adaptation (UDA) emerges, which shows promising results on object detection [9]–[22] and semantic segmentation [23]–[45], aiming to reduce the domain shifts between the source and the target domains.

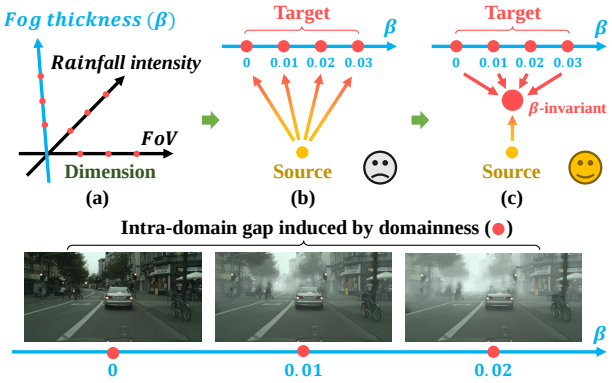


Fig. 1: (a) Previous UDA methods do not leverage explicit prior knowledge to perform domain adaptation on a demand-specific dimension, and (b) they can not generalize well to a target domain with different unknown domainness. (c) Our method narrows the intra-domain gap induced by different domainness. Different domainness indicate the different numerical magnitudes of a specific domain dimension.

The domain shifts may span different dimensions such as fog, rainfall, Field of View (FoV), etc. A practical scenario is to align the source and the target domain in a demanded specific dimension, *e.g.,* from sunny images to foggy images. However, existing methods can hardly handle such cases elegantly. This is mainly because they do not consider any explicit prior knowledge about the specific domain shifts. As a result, the model will lack a clear target dimension and will be optimized without the aforementioned prior knowledge. This under-constrained training process limits the performance when adapting the model to a specific dimension. Meanwhile, the intra-domain gaps, which commonly exist among the domains in the same dimension but with different domainness, are taken out of consideration in previous research. As shown in Fig. 1, different *domainness* values indicate different numerical magnitudes of a specific domain dimension. Narrowing such intra-domain gaps is crucial when adapting to a specific dimension.

In this paper, we refer to the above explained problem as Specific Domain Adaptation (SDA), a realistic and practical setting for domain adaptation. It targets to generalize a model across a specific domain dimension, *e.g.,* different FoVs (FoV dimension), and the model can be broadly applied in real-

Manuscript received xx xx, 2021. revised xx xx 2021.

Q. Zhou, Q. Gu, F. Zheng are with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: {zhouqianyu, miemie, zyfeng97}@sjtu.edu.cn).

J. Pang is with the Multimedia Laboratory, The Chinese University of Hong Kong, China (e-mail: pangjiangmiao@gmail.com).

G. Cheng and J. Shi are with SenseTime Research, Beijing, China (e-mail:guangliangcheng2014@gmail.com, shijianping@sensetime.com).

X. Lu is with the School of Information Technology, Deakin University, Victoria 3216, Australia (e-mail: xuequan.lu@deakin.edu.au).

L. Ma is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the School of Computer Science and Technology, East China Normal University, Shanghai 200062, China (e-mail: ma-lz@cs.sjtu.edu.cn).

* Q. Zhou, and Q. Gu contribute equally.

world applications. For example, in autonomous driving, the models trained on the sunny days should have the ability to generalize to the specific rainy or foggy scenarios.

To address the above SDA, we propose an innovative method, referred to as Self-Adversarial Disentangling (SAD). It tackles the problem by disentangling the latent representations into domainness-invariant features and domainness-specific features in a specific dimension. In contrast to domain-invariant feature that involves some noise factors, domainness-invariant feature means the feature is irrelevant to the domainness magnitude in the target domain. The advantage of transferring domainness-invariant features is that we can capture the generalizations across different domainness to narrow the intra-domain gaps, which is on a more fine-grained level than directly transferring domain-invariant features.

Our framework consists of two key components, *i.e.,* Domainness Creator (DC) and Self-Adversarial Regularizer (SAR), for domainness generation and disentanglement, respectively. According to the given domain shift, we firstly enrich the source domain with DC. It not only diversifies the source domain but also provides additional supervisory signals for the following feature disentangling. Guided by the domainness, we design the SAR, and introduce a domainness-specific loss and a domainness-invariant loss for SAR to jointly supervise the disentangling of the latent representations into domainness-specific and -invariant features. With the domainness-specific loss, our SAR can classify the predicted domainness with supervisory signals from DC. Penalized by the domainness-invariant loss, our SAR can fully learn domainness-invariant representations. Thus, we can mitigate the intra-domain gap induced by different domainness. To sum up, our SAD framework works in a disentangling sense, which enables the model to learn domainness-invariant feature in an adversarial manner, *i.e.,* two opposite loss functions.

Our method is applicable and flexible in most real-world cases. We verified the proposed method under various domain dimensions, including cross-fog (Cityscapes [46] to Foggy Cityscapes [47], Cityscapes [46] to RTTS [48], Cityscapes [46] to Foggy Zurich++ [47], [49]), cross-rain (Cityscapes [46] to RainCityscapes [50]) and cross-FoV adaptation (Virtual KITTI [51] to CKITTI [46], [52]). The target domain has either single or multiple domainness values. Extensive experiments prove the impressive generalization abilities of our method. Without bells and whistles, our method yields remarkable improvements over existing methods in both object detection and semantic segmentation. In particular, we achieve $3.4\% \sim 6.4\%$ gains on synthetic datasets and improvements of up to 2.6% on real datasets. We achieve 45.2% mAP on the widely-used benchmark of Cityscapes [46] to Foggy Cityscapes [47].

The contributions of this paper are summarized as follows. 1) We study the problem of specific domain adaptation (SDA), a realistic and practical setting for domain adaptation. We propose a novel self-adversarial disentangling framework by leveraging the explicit prior domain knowledge to learn the domainness-invariant features. 2) We present a domainness creator for specifically enriching the source domain and providing explicit supervisory signals. 3) We design a self-adversarial regularizer to mitigate the intra-domain gaps. We also introduce one domainness-specific loss and a domainness-invariant loss to facilitate the training. 4) We conduct comprehensive experiments to demonstrate the effectiveness of our method on both object detection and semantic segmentation. It is simple to integrate our method into any existing approaches as a plug-and-play framework which does not introduce any extra costs during the inference phase.

II. Related Work

Unsupervised Domain Adaptation. UDA aims to generalize the model learned from the labeled source domain to another unlabeled target domain. In the field of UDA, a group of approaches has shown promising results in object detection [9]–[22] and semantic segmentation [23]–[45], [53], [54]. The current mainstream approaches of these two tasks include adversarial learning [10]–[12], [26], [29], [55], self-training [8], [41], [42], [56] and self-ensembling [39], [54], [57]–[63]. Despite the gratifying progress, little attention has been paid to perform domain adaptation in a specifically demanded dimension by introducing any explicit prior knowledge about the domain shifts except [19]. Prior DA [19] is the only work that builds on a similar motivation with us by using the weather-specific prior knowledge obtained from the image formation. However, it designed a prior-adversarial loss and acts in a completely different manner from ours. Prior DA [19] only explored the weather prior on the cross-fog and cross-rain scenarios. We follow the same setting as [19] by knowing the domain dimension in advance, which is fully fair in experimental comparison.

Domain Diversification. Domain Diversification (DD) aims to diversify the source domain to various distinctive domains with random augmentation. Kim *et al.* [64] designed a DD-MRL method by using GAN [65] to diversify the source domain. Similarly, DRPC [66] and LTIR [28] proposed to diversify the texture of the source images and to learn texture-invariant representations. *Our method differs from these methods in several aspects.* Firstly, they require large computation costs and cannot be trained end-to-end during the adaptation procedure. Instead, our method is light-weighted and online with a transformation algorithm in DC. Secondly, the GAN-based approaches tend to produce artifacts for urban-scene datasets, leading to severe semantic inconsistency. In contrast, we do not use any feature interpolation operation in the reconstruction and merely use a simple yet very effective parameter modeling.

Disentangled Learning. Disentangled learning has been widely studied in other communities, *e.g.,* image translation [67], [68], few-shot learning [69], [70]. A few works have recently extended it into domain adaptation by disentangling the latent representations into domain-specific and domain-invariant features to realize effective domain alignment. Liu *et al.* proposed a model of cross-domain representation disentanglement (CDRD) [71] based on the GAN [65] framework. Chang *et al.* designed a domain invariant structure extraction (DISE) framework [31] to disentangle the latent encodings into the domain-invariant structure and domain-specific texture representations for domain-adaptive semantic segmentation.

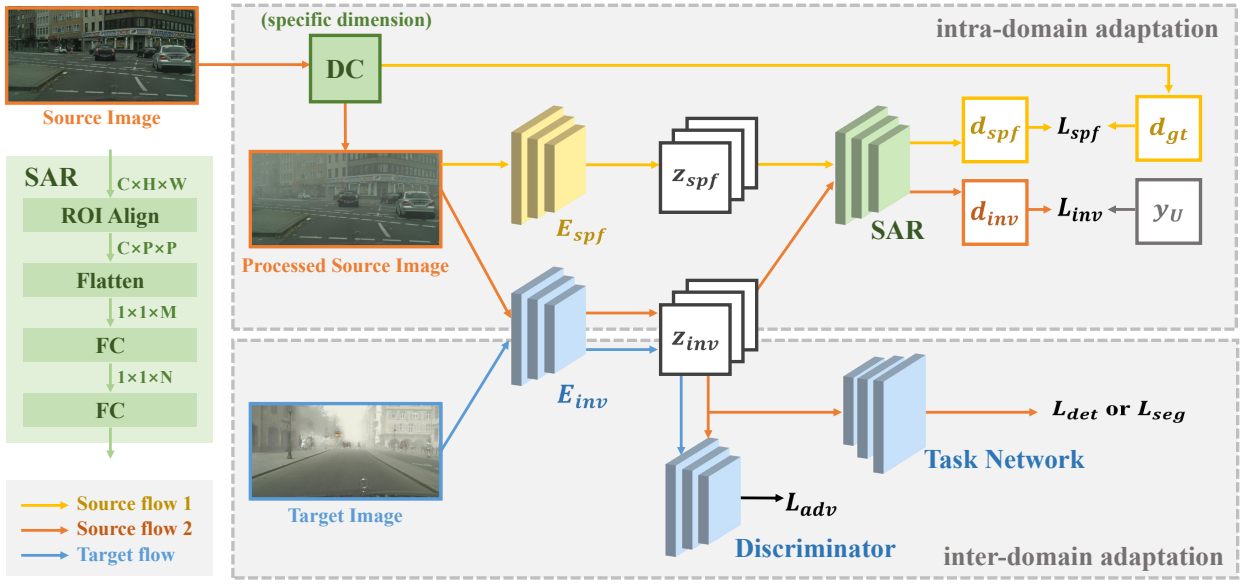


Fig. 2: Overview of the proposed Self-Adversarial Disentangling (SAD). Our Domainness Creator (DC) not only generates a diversified source image with random domainness, but also provides additional supervisory signals d_{gt} for guiding the self-adversarial learning. The encoder E_{spf} and E_{inv} are to extract the domainness-specific representations z_{spf} and the domainness-invariant representations z_{inv}, respectively. With the guidance of the generated domainness, E_{spf}, E_{inv} and SAR (Self-Adversarial Regularizer) work in an adversarial manner, *i.e.,* two opposite loss functions, to disentangle the latent representations into z_{spf} and z_{inv}.

Nevertheless, these methods can hardly capture the generalizations across different domainness within the same target domain to narrow down the intra-domain gap.

Multi Domain-invariant Representation Learning. The most relevant work to ours in the filed of UDA is multi domain-invariant representation learning (MRL) [64]. MRL applies a multi-domain discriminator to learn indistinguishable features among different domains. Similarly, Wang *et al.* [72] proposed a multi-domain discriminator that models the encoding-conditioned domain index distribution to tackle the continuously indexed domain adaptation (CIDA). *Our method is quite different from these methods*. Firstly, these approaches do not leverage the prior domain knowledge in a specific dimension to facilitate the multi-domain representation learning. In contrast, we model a specific dimension in domain shift and leverage the generated domainness as supervisory signals to guide the feature learning. Secondly, the multi-domain discriminators are not actually reconstructing the domainness values, and they neglect the intra-domain gaps within the target domain induced by different domainness. Instead, guided by the domainness-invariant and domainness-specific loss functions, our SAR works in a completely different manner to narrow the intra-domain gaps.

III. Methodology

We focus on the problem of Specific Domain Adaptation (SDA) in both object detection and semantic segmentation, where we have access to the source data X_S with labels Y_S and the target data X_T without labels. Fig. 2 shows the overview of our framework. Our core idea is to disentangle the latent representation into domainness-invariant feature and domainness-specific feature in a specific dimension. The target domain has either single or multiple domainness values.

A. Domainness Creator

We design Domainness Creator (DC) as a transformation algorithm for images. DC receives a source image X_S as the input and outputs a processed image $\tilde{X}_S$ by adding a random domainness in a specific dimension. Meanwhile, DC provides a supervisory signal, *i.e.,*, the label of the domainness value d_{gt}, for guiding the self-adversarial learning. Due to the variations of domainness values enabled by DC, a model trained on the domainness-diversified dataset will be able to learn the domainness-invariant representations for feature alignment. d_{gt} is a number, *e.g.,* FoV_x is $40°$. FoV_x denotes the FoV in the x axis.

Example of DC in FoV dimension. Taking FoV as an example, we show the process of FoV transformation given a selected FoV_x in Fig. 3, where O is the optical center of the camera and F is the focal point. OF denotes the focal length. MN and PQ represent the original width and the new width before and after the transformation:

$$\tilde{X}_S = DC(X_S), FOV_x = \angle MFN \rightarrow \angle PFQ \qquad (1)$$

where FoV_x is reduced from $\angle MFN$ to $\angle PFQ$ during the process and the domainness label is denoted as $d_{gt} = \angle PFQ$. If the dimension is fog thickness for DC, we follow the algorithms in [47] to diversify the source image.

Remark 1: *The intuitions between DC and data augmentation are totally different.* Data augmentation only diversifies the source images, while the proposed DC provides supervision

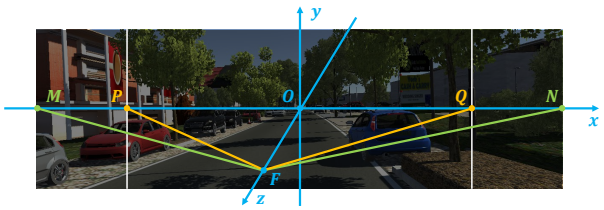


Fig. 3: The FoV_x transform. O is the optical center of the camera and F is the focal point. OF is the focal length, MN and PQ represent the original width and the new width before and after the transformation, respectively. FoV_x is reduced from $\angle MFN$ to $\angle PFQ$ after the process.

signals to guide the training of SAR, which is the critical part of the method. DC indeed helps the disentanglement of learning the domainness-invariant representations. We also prove that the two components are complementary in the experiment part.

Remark 2: *Difference from existing GAN-based domain diversification.* Compared to previous domain diversification [64] and domain randomization [66] approaches, we do not use GAN-based architecture to produce translated images in our implementations. The main reasons are reflected in two aspects. Firstly, these methods require large computation resources, especially for the urban-scene images, and they cannot be trained in an end-to-end manner. Secondly, the reconstruction of the feature encodings will inevitably lead to pixel-wise distortions and semantic inconsistency. In comparison, our transformation in DC is online and allows end-to-end training, because we do not use any feature interpolation operation and just use a simple yet effective mathematically modeled transformation.

B. Self-Adversarial Regularizer

Guided by the generated domainness, SAR is designed to disentangle the latent representations into the domainness-specific feature z_{spf} and the domainness-invariant feature z_{inv}, in order to mitigate the intra-domain gap. E_{spf} and E_{inv} denote the domainness-specific encoder and domainness-invariant encoder, respectively. The dimensions of z_{spf} and z_{inv} are both $C * H * W$, where C is 19/11 for segmentation, and 512/1024 for detection, respectively.

Intra-domain adaptation. As shown in Fig. 2, the processed image $\tilde{X}_S$ is fed into the encoder E_{spf} and E_{inv} to get the latent feature map z_{spf} and z_{inv}. Either z_{spf} or z_{inv} is forwarded into SAR to get the domainness value d_{spf} and d_{inv} for once. SAR is supervised by the designed domainness-specific loss $\mathcal{L}_{spf}$ and domainness-invariant loss $\mathcal{L}_{inv}$ together (see below for the design of these two losses). With the former loss $\mathcal{L}_{spf}$, our SAR could classify the predicted domainness d_{spf} with supervisory signals d_{gt} from DC. Penalized by the latter loss $\mathcal{L}_{inv}$, our SAR could fully learn domainness-invariant representations, thus mitigating the intra-domain gap induced by different domainness. In essence, the encoders and SAR are *complementary* and work in an *adversarial* manner (*i.e.,* two opposite losses) to perform the specific domain adaptation. We illustrate the network details and the two loss fuctions below.

Network architecture of SAR. Note that we use the same SAR architecture for both detection and segmentation tasks. Our SAR only takes one feature map z_{spf} or z_{inv} at a time as input. After that, we downsample the whole feature map to predict domainness value, and then flatten the downsampled feature map. Then after two fully-connected layers with a relu activation, we get the domainness value d_{spf} or d_{inv}, as shown in Fig. 2. In practice, we use ROI Align [73] to downsample the *whole* feature map to predict domainness value. We discretize the continuous domainness values into N numbers (representing N ranges) for better experimental results. d_{spf}, d_{inv} are one-hot vectors with N dimensions. y_U is a N dimensional vector of the uniform distribution.

Domainness-specific loss. On one hand, with the generated domainness d_{gt} as a supervisory signal, SAR needs to enhance its discriminativity for classifying the diversified images with different domainness more accurately. We define the domainness-specific loss $\mathcal{L}_{spf}$ as a cross-entropy loss for optimizing the features from the encoder E_{spf}:

$$\mathcal{L}_{spf} = -\sum_{i=1}^{N} d_{gt}^{i} log(d_{spf}^{i}), \tag{2}$$

where d_{gt} is now used as the one-hot vector of generated domainness and d_{spf} is the predicted domainness value of SAR.

Domainness-invariant loss. On the other hand, SAR needs to maximize the discrepancy between the domainness-invariant feature z_{inv} and the domainness-specific feature z_{spf}. We define the domainness-invariant loss $\mathcal{L}_{inv}$ as the KL-divergence between the predicted domainness d_{inv} and a uniform distribution:

$$\begin{aligned} z_{inv} &\sim E_{inv}\left(\tilde{X}_S\right) = q_S\left(d_{inv} \mid \tilde{X}_S\right), \\ \mathcal{L}_{inv} &= KL\left(q_S\left(d_{inv} \mid \tilde{X}_S\right) \| p\left(x\right)\right), \end{aligned} \tag{3}$$

where x is sampled from a uniform distribution y_U, $p(x)$ is the probability of x, and q_S denotes the distribution of domainness d_{inv}.

By jointly minimizing the domainness-invariant loss $\mathcal{L}_{inv}$ and the domainness-specific loss $\mathcal{L}_{spf}$ in two inverse directions, SAR can fully learn the domainness-invariant features which capture the generalizations across different domainness, thus narrowing the intra-domain gap.

Remark 1: *Whether the parameters of E_{spf} and E_{inv} are shared or not.* E_{spf} and E_{inv} are two encoders that use the same architecture but do not share the weights, because they are penalized by different loss functions. The former is penalized by L_{spf}, and the latter is under the guidance of L_{inv}, L_{adv} (adversarial loss, Eq. (4)) and L_{task} (task loss, Eq. (6)).

Remark 2: *Comparing with GAN architecture.* Existing GAN-based architectures utilized the multi-domain discriminators [64], [72] to distinguish the domainness (they called domain index in their work). In the adversarial framework, these discriminators are not actually predicting the domainness d_{inv}, but making the latent encodings z_{inv} unable to predict

d_{inv}. Due to the fact that it is trained in an adversarial way, the encoder will transform the input X before outputting encoding z_{inv}, thereby removing the information related to domainness d_{inv}. However, the encoder can not fully learn the domainness-invariant feature due to the lack of prior knowledge about the domain shifts. *In comparison, our proposed framework acts in a completely different manner.* Firstly, we use two separate encoders E_{spf} and E_{inv}, the former for extracting the domainness-specific feature z_{spf} and the latter for extracting the domainness-invariant features z_{inv}. Secondly, with the guidance of the generated domainness d_{gt} as supervisory signals, our SAR is truly reconstructing the domainness, aiming to distinguish the domainness accurately; Thirdly, our SAD framework works in two opposite directions in a disentangling sense, which enables the model to learn domainness-invariant features to alleviate the intra-domain gap.

C. End-to-End Training and Inference

In this section, we will briefly introduce the inter-domain adaptation, the task loss and formulate an overall loss function for end-to-end training. Then we will explain the inference phase.

Inter-domain adaptation. Without loss of generality, we employ an adversarial framework [74] for the inter-domain adaptation. As shown in Fig. 2, the processed source images $\tilde{X}_S$ and X_T are fed into the encoder E_{inv}. Then, E_{inv} is encouraged to learn z_{inv}. The latent encodings should confuse a domain discriminator D in distinguishing the features extracted between the source and target domains. This is achieved by min-maximizing an adversarial loss:

$$\begin{aligned}\mathcal{L}_{adv} = &- \mathbb{E}_{\mathbf{x}\sim p(\tilde{\mathbf{x}}_S)}[\log(D(E_{inv}(\mathbf{x})))] \\ &- \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}_T)}[\log(1 - D(E_{inv}(\mathbf{x})))],\end{aligned} \tag{4}$$

Task loss. In this work, taking Faster-RCNN [3] as an example, we use Region Proposal Network (RPN) to generate Region of Interests (RoIs). It then localizes and classifies the regions to obtain semantic labels and locations. The task network is optimized with a multi-task loss function:

$$\mathcal{L}_{task} = \mathcal{L}_{rpn} + \lambda_{cls}\mathcal{L}_{cls} + \lambda_{reg}\mathcal{L}_{reg}, \tag{5}$$

where the RPN loss $\mathcal{L}_{rpn}$, classification loss $\mathcal{L}_{cls}$ and regression loss $\mathcal{L}_{reg}$ remain the same as [3]. The loss weights λ_{cls} and λ_{reg} are set to 1.0 by default.

Total loss. During training, all the models are jointly trained with the backbone in an end-to-end manner. The total loss L_{total} is the weighted sum of the aforementioned loss functions:

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda_{adv}\mathcal{L}_{adv} + \mathcal{L}_{inv} + \lambda_{spf}\mathcal{L}_{spf}, \tag{6}$$

where λ_{adv} and λ_{spf} are the weighting coefficients for the loss $\mathcal{L}_{adv}$ and $\mathcal{L}_{spf}$, respectively. We use the original weighting ratio in [9], [10], [13], [16], [26], [29], [30] to balance $\mathcal{L}_{task}$ and $\mathcal{L}_{adv}$.

Inference phase. In the inference phase, we only need a domainness-invariant encoder E_{inv} with a task network T to make predictions. In other words, all other modules including DC, SAR and E_{spf} are removed in the inference stage, leading to no extra costs in prediction. Besides, our method can be plugged into various existing cross-domain detection/segmentation methods. Thus, our framework is flexible and generalizable, and it does not depend on specific UDA frameworks for feature alignment.

IV. Experiments

In this section, we firstly evaluate our framework on object detection under various domain dimensions, including cross-fog adaptation, cross-rain adaptation and cross-FoV adaptation. In addition, we extend our method to the semantic segmentation to verify its scalability and applicability. Finally, we conduct ablation studies to validate each component of our method. Our method is applicable and flexible in most real-world cases, and we proved it with thorough experiments.

A. Datasets

Cityscapes → Foggy Cityscapes. This is a widely-used benchmark for cross-domain object detection. Cityscapes [46] is a dataset focused on autonomous driving, which consists of 2,975 images in the training set, and 500 images in the validation set. Foggy Cityscapes [47] is a synthetic foggy dataset which simulates fog on real scenes. The annotations and data split in Foggy Cityscapes are inherited from Cityscapes.

Cityscapes → RTTS. RTTS [48] is the largest available dataset for object detection under real-world hazy conditions. It contains 4,807 unannotated and 4,322 annotated real-world hazy images covering most traffic and driving scenarios with 7 kinds of fogs.

Cityscapes → Foggy Zurich++. Foggy Zurich++ is a real-world dataset collected in foggy-weather conditions for segmentation. We use all the unannotated 3,768 images of Foggy Zurich [47] as the training set and mix the validation set of Foggy Driving [49] and Foggy Zurich [47]. Following [46], Foggy Zurich++ is labeled with 19 classes.

Cityscapes → RainCityscapes. RainCityscapes [50] renders Cityscapes images with synthetic rain. Each clear image is rendered with 12 types of rain patterns, including 4 types of drop sizes which we use as our domainness. The annotations are the same as those of Cityscapes. We use this benchmark in cross-domain detection.

VKITTI → CKITTI. We use this benchmark in both detection and segmentation. Virtual KITTI [51] is a photo-realistic synthetic dataset, which contains 21,260 images. It is designed to mimic the conditions of KITTI dataset and has similar scene layouts, camera viewpoints and image resolution to KITTI dataset. CKITTI is a real-world dataset depicting several urban driving scenarios with 5 different kinds of FoVs, which is a mixed dataset of Cityscapes [46] and KITTI [52]. We use the 10,456 images as the training set and 700 images as the validation set.

B. Implementation Details

Object detection. In our implementation, we follow the training protocol [9], [10], [13], [16] of the Faster-RCNN network. We resize the images of both the source and target

TABLE I: Cross-fog (a,b) and cross-FoV (c) adaptation of object detection.

(a) Cityscapes to Foggy Cityscapes (single-domainness).

Methods	Venue	person	rider	car	truck	bus	train	motor	bicycle	mAP
Source-Only [3]	NeurIPS'15	26.9	38.2	35.6	18.3	32.4	9.6	25.8	28.6	26.9
DA-Faster [9]	CVPR'18	29.2	40.4	43.4	19.7	38.3	28.5	23.7	32.7	32.0
SCDA [12]	CVPR'19	33.5	38.0	48.5	26.5	39.0	23.3	28.0	33.6	33.8
DD-MRL [64]	CVPR'19	31.8	40.5	51.0	20.9	41.8	34.3	26.6	32.4	34.9
MAF [11]	ICCV'19	28.2	39.5	43.9	23.8	39.9	33.3	29.2	33.9	34.0
ART-PSA [21]	CVPR'20	34.0	46.9	52.1	30.8	43.2	29.9	34.7	37.4	38.6
ICR-CCR [15]	CVPR'20	32.9	43.8	49.2	27.2	45.1	36.4	30.3	34.6	37.4
CST [20]	ECCV'20	32.7	44.4	50.1	21.7	45.6	25.4	30.1	36.8	35.9
ATF [18]	ECCV'20	34.6	47.0	50.0	23.7	43.3	38.7	33.4	38.8	38.7
Prior DA [19]	ECCV'20	36.4	**47.3**	51.7	22.8	47.6	34.1	36.0	38.7	39.3
GPA [13]	CVPR'20	32.9	46.7	54.1	24.7	45.7	41.1	32.4	38.7	39.5
Ours (with GPA [13])	-	**38.3**	47.2	**58.8**	**34.9**	**57.7**	**48.3**	**35.7**	**42.0**	**45.2**

(b) Cityscapes to RTTS (multi-domainness).

Methods	car	bus	person	motor	bicycle	mAP
Source-Only	39.8	11.7	46.6	19.0	37.0	30.9
DCPDN [75]	39.5	12.9	48.7	19.7	37.5	31.6
Grid-Dehaze [76]	25.4	10.9	29.7	13.0	21.4	20.0
DA-Faster [9]	43.7	16.0	42.5	18.3	32.8	30.7
SWDA [10]	44.2	16.6	40.1	23.2	41.3	33.1
Ours (with [9])	45.0	15.9	42.0	22.2	38.4	32.7
Ours (with [10])	**47.0**	**16.6**	41.5	**27.2**	**43.2**	**35.1**

(c) Virtual KITTI to CKITTI (multi-domainness)

Comparisons	Car AP	Gain
DA-Faster [9]	45.1	**2.6**
Ours (with [9])	**47.7**	
SWDA [10]	49.0	**1.7**
Ours (with [10])	**50.7**	
SCL [16]	49.5	**1.8**
Ours (with [16])	**51.3**	

TABLE II: Cross-rain adaptation of object detection from Cityscapes to RainCityscapes (multi-domainness).

Methods	person	rider	car	truck	bus	motor	bicycle	mAP	Gain
DA-Faster [9]	22.9	55.2	43.4	3.9	58.8	15.2	30.0	32.8	**6.4**
Ours (with DA-Faster [9])	**26.3**	**60.1**	**52.6**	**13.0**	**60.3**	**27.0**	**34.9**	**39.2**	
SWDA [10]	23.8	52.1	46.4	9.6	68.2	16.0	32.8	35.6	**3.4**
Ours (with SWDA [10])	**25.9**	**56.0**	**52.5**	8.1	**56.0**	**29.4**	**33.1**	**39.0**	
SCL [16]	27.0	57.9	50.3	10.0	67.9	13.9	33.9	37.3	**4.2**
Ours (with SCL [16])	**29.3**	**61.0**	**52.7**	**19.2**	**68.2**	**26.2**	**34.1**	**41.5**	

domain to 600-pixel height in all experiments as suggested by [9], [10], [16]. Following the aforementioned papers, we use the VGG16 [77] model pre-trained on ImageNet [78] as a backbone of DA-Faster [9], SWDA [10] and SCL [16], and the ResNet50 [79] as the backbone of GPA [13]. We set the learning rate to 0.001 for the first 50k iterations and 0.0001 for the rest iterations. The other parameters are set by following the original papers [9], [10], [13], [16].

Semantic segmentation. Following common UDA protocols [26], [29], [30], we employ the DeepLab-v2 [5] with ResNet 101 backbone [79] in our implementation. The backbone is pre-trained on ImageNet [78]. We reproduce the famous AdaptSegNet [26], CLAN [29] and SIM [30] as our baselines. For our DeepLab-v2 network, we use Adam as the optimizer. The initial learning rate is 2.5×10^{-4}, which is then decreased using polynomial decay with an exponent of 0.9.

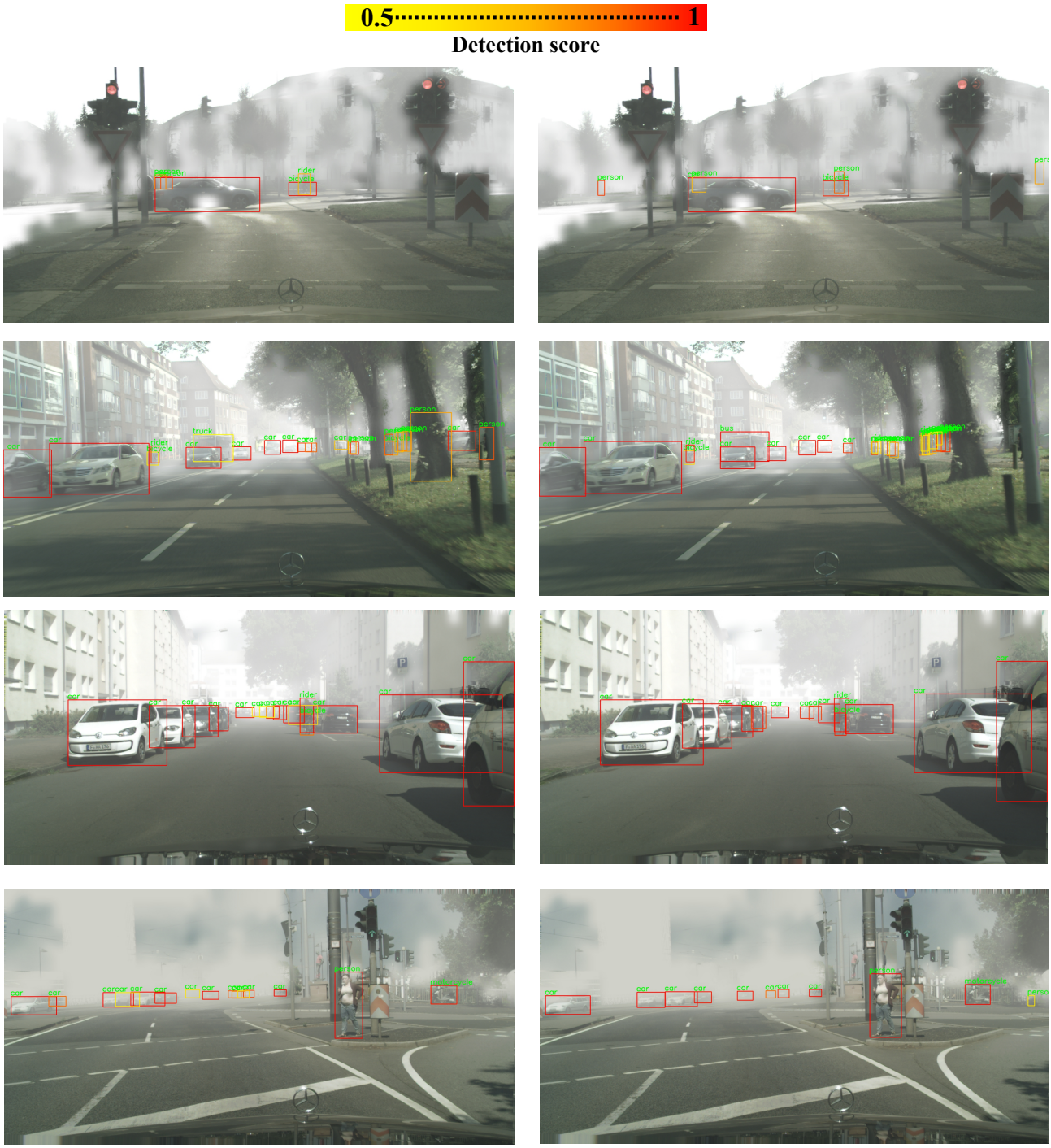
C. Domain Adaptation for Object Detection

In this section, we present the results in three dimensions, *i.e.,* cross-fog, cross-rain and cross-FoV adaptation, to show the effectiveness of our approach. We achieve $3.4\% \sim 6.4\%$ gains on synthetic datasets and improvements of up to 2.6% on real datasets.

Cross-fog adaptation. To validate the generalization capability on the cross-fog adaptation, we perform two experiments, where the target domain includes single and multiple domainness values, respectively.

Single domainness within the target domain: In this experiment, we adapt from Cityscapes [46] to Foggy-Cityscapes [47]. Table I (a) presents the comparison results with the state-of-the-art cross-domain detection methods on eight categories. Source-only indicates the baseline Faster RCNN [3] is trained with only source domain data. From the table, we can observe that our method (with GPA [13]) could outperform the state-of-the-arts by 5.7%. The published Prior-DA [19] builds on a similar motivation as ours by using the weather-specific prior knowledge obtained from the image formation. It designed a prior-adversarial loss and acts in a completely different manner. Also, [19] knows the dimension in advance, which fully proves our fairness of this setting. Our method outperforms Prior-DA [19] by 5.9%.

Taking a closer look at per-category performance in Table I (a), our approach achieves the highest AP on those cate-



(a) GPA **(b) Ours (with GPA)**

Fig. 4: Qualitative results of cross-domain object detection in the Cityscapes [46] → Foggy Cityscapes [47] set-up. The two columns plot (a) the predictions of GPA [13] baseline, (b) the predictions of Ours (with GPA [13]). The bounding boxes are colored based on the detector confidence using the shown color map. As we can see from the results, the proposed method is able to produce high confidence predictions and is able to detect more objects in the images.

TABLE III: Cross-FoV adaptation of semantic segmentation from Virtual KITTI to CKITTI (multi-domainness).

Method	road	building	pole	light	sign	vegetation	terrain	sky	car	truck	guard rail	**mIoU$_{11}$**	**Gain**
AdaptSegNet [26]	88.0	80.6	11.1	17.4	28.4	80.3	29.2	85.2	82.1	29.7	27.5	50.8	**1.8**
Ours (with AdaptSegNet [26])	**88.4**	**81.0**	9.7	**18.9**	**30.5**	**80.9**	**39.1**	**86.2**	**83.6**	**32.6**	**27.5**	**52.6**	
CLAN [29]	88.2	80.0	6.0	17.9	26.7	79.3	36.1	85.7	82.4	28.5	12.3	49.4	**1.1**
Ours (with CLAN [29])	88.1	79.9	**9.9**	**19.6**	25.3	**80.2**	**38.5**	**85.9**	**82.5**	**29.2**	**16.4**	**50.5**	
SIM [30]	87.3	81.2	16.3	16.1	28.3	81.6	37.6	87.2	82.6	29.3	18.3	51.4	**1.8**
Ours (with SIM [30])	86.7	**81.9**	15.7	**17.7**	**31.7**	**82.3**	**48.2**	**86.6**	81.9	**32.3**	**20.4**	**53.2**	

TABLE IV: Cross-Fog adaptation of semantic segmentation from Cityscapes to Foggy Zurich++ (multi-domainness).

Method	mIoU	Gain
AdaptSegNet [26]	29.4	**5.8**
Ours (with AdaptSegNet [26])	**35.2**	
CLAN [29]	26.8	**4.7**
Ours (with CLAN [29])	**31.5**	
SIM [30]	27.0	**4.1**
Ours (with SIM [30])	**31.1**	

gories. This phenomenon illustrates the effectiveness of Self-Adversarial Disentangling among different classes during the cross-domain detection.

Multiple domainness within the target domain: In this experiment, we adapt from Cityscapes [46] to RTTS dataset [48]. Multi-domainness means there exist 7 kinds of fogs in RTTS dataset. The comparison results with the state-of-the-arts are reported in Table I (b). As for the image dehazing approaches which dehaze the target domain and then trasfer the domain knowledge, DCPDN [75] improves the Faster RCNN performance by 1%. However, Grid-Dehaze [76] does not help the Faster RCNN baseline and results in even worse performance. Table I (b) shows that our method can effectively boost the performance by integrating it into DA-Faster RCNN [9] and SWDA [10]. We can successfully boost the mAP by 2.0% and 2.0%, respectively. The benefits of our approach lie in two aspects: (1) our method can be easily adopted as a plug-and-play framework which enables end-to-end training and no extra costs during the inference time. (2) Our approach can not only address the single domainness problem but also tackle more complicated scenarios where multiple domainness exist in the target domain.

Cross-FoV adaptation. To validate the generalization capability of the proposed method, we also conduct an experiment on the FoV dimension adapting from Virtual KITTI [51] to CKITTI [46], [52]. The adaptation results are reported in Table I (c). Despite the 5 different FoVs in the dataset, our method can always achieve a certain improvement. By plugging into the current state-of-the-art methods, *i.e.,* DA-Faster [9], SWDA [10], SCL [16], our method brings 2.6%, 1.7% and 1.8% increase, respectively.

Cross-rain adaptation. We conduct experiments from Cityscapes [46] to RainCityscapes [50]. Table II shows the results of adapting the model between different rain scenarios. We reproduce DA-Faster RCNN [9], SWDA [10] and SCL [16] in the same setting. We can see that our method significantly improves the mAP by 6.4%, 3.4% and 4.2% through integrating it into the existing UDA methods.

D. Domain Adaptation for Semantic Segmentation

In addition to the above experiments on cross-domain object detection, we also conduct experiments on cross-domain semantic segmentation, to show the scability of our method. In specific, we conduct the cross-FoV adaptation and cross-fog adaptation on semantic segmentation.

Cross-fog adaptation. In this experiment, we adapt from Cityscapes [46] to Foggy Zurich++ [47], [49] to perform the cross-fog adaptation, where multiple thickness of fogs exist in the target domain. As shown in Table IV, our method outperforms the state-of-the-art methods [26], [29], [30] by 5.8%, 4.7% and 4.1%, respectively.

Our method can handle the cases where a domainness value is never seen in the training stage and we have verified it with experiments. As shown in Table IV, the Foggy Zurich++ has the real fog rather than the synthetic fog, which means the domainness in the validation set is unknown and does not appear in the training set. Our method works well on this dataset, which proves its generalization ability.

Cross-FoV adaptation. In this experiment, we perform the specific domain adaptation given the FoV gap. We choose Virtual KITTI [51] as the source domain and CKITTI [46], [52] as the target domain. The comparison results are listed in Table III. Compared with the AdaptSegNet [26], CLAN [29] and SIM [30], our method respectively yields an increase of 1.8%, 1.1% and 1.8%, which indicates the effectiveness of the proposed SAD in the semantic segmentation task and shows its good scalability.

E. Ablation Studies

In this section, we perform ablation experiments to investigate the effect of each component and provide more insights of our method.

Comparisons to the related work. Table V shows the comparisons to the relevant work [64], [72] from Cityscapes [46] to Foggy Zurich++ [47], [49] under the same baseline. When using MRL [64] or CIDA [72] as the adaptor, it merely

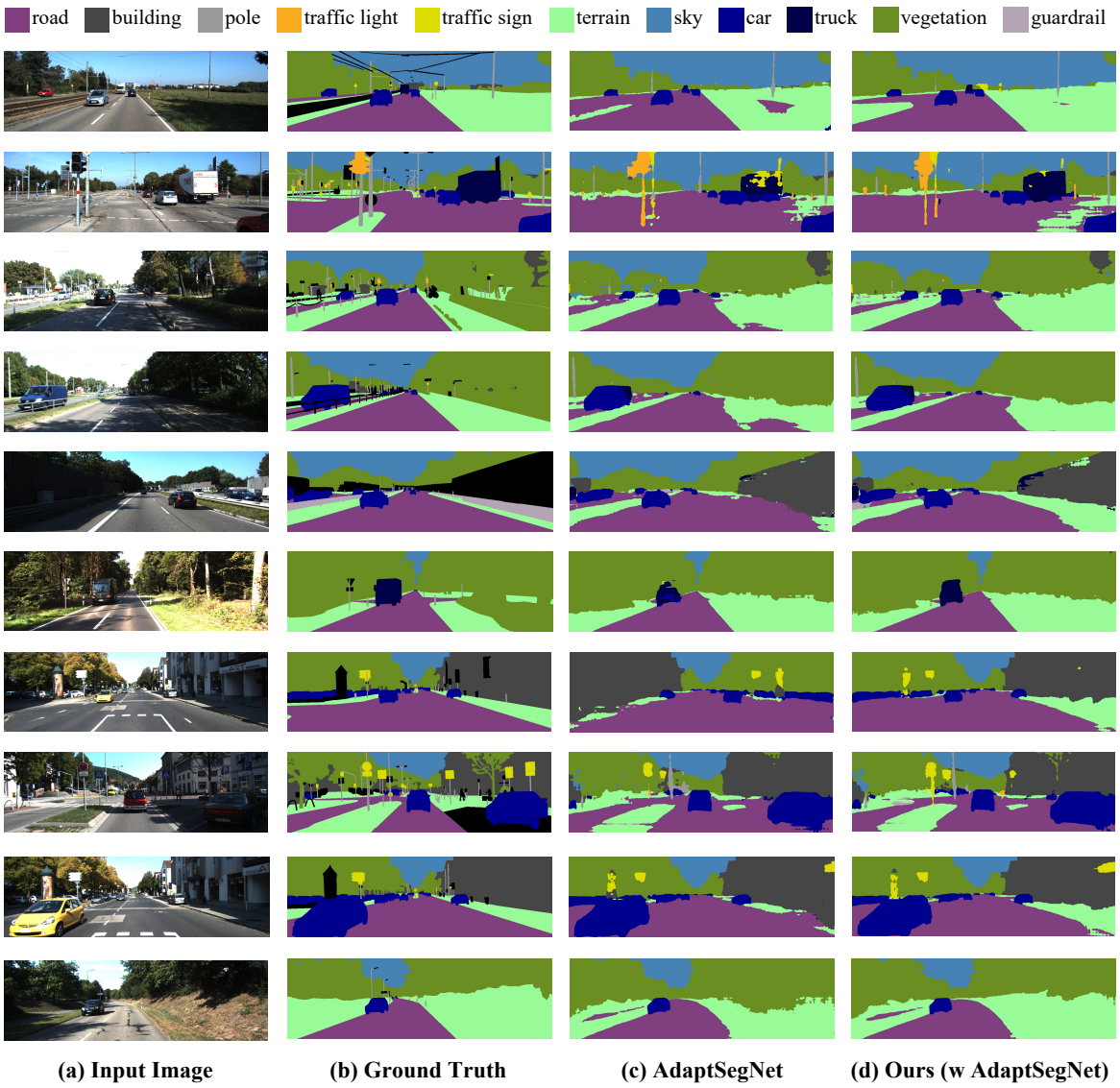


Fig. 5: Qualitative results of cross-domain semantic segmentation in Virtual KITTI [51] → CKITTI [46], [52] (11 classes) set-up. The four columns plot (a) RGB input images, (b) ground truth, (c) the predictions of AdaptsegNet [26] baseline and (d) the predictions of Ours (with AdaptsegNet [26]).

TABLE V: Ablation on Cityscapes to Foggy Zurich++.

Baseline	Diversificator	Adaptor	mIoU	Gain
[26]	-	-	29.4	-
	-	MRL [64]	29.9	0.5
	-	CIDA [72]	30.0	0.6
	-	Ours (SAR)	**30.8**	**1.4**
	DD [64]	-	32.3	2.9
	Ours (DC)	-	**33.5**	**4.1**
	DD [64]	MRL [64]	33.7	4.3
	Ours (DC)	MRL [64]	34.0	4.6
	Ours (DC)	CIDA [72]	34.2	4.8
	Ours (DC)	Ours (SAR)	**35.2**	**5.8**

achieves a limited improvement of 0.5% or 0.6%. In contrast, SAR contributes to the performance gain of 1.4%. The diversificator, *e.g.,* GAN-based DD [64] only brings 2.9% gain over the baseline, while our DC boosts the baseline by 4.1%.

The main reasons are twofold. (1) Previous GAN-based methods [64], [72] do not utilize supervisory signals d_{gt} from DC to fully learn the domainness-invariant feature. (2) They neglect the intra-domain gap induced by different domainness. Instead, our method not only leverages the prior supervisory signals but also mitigates the intra-domain gap across different domainness. Incorporating DC and SAR into the same framework boosts the mIoU by 5.8% over the baseline. This confirms the effectiveness of our proposed DC and SAR, and addresses the aforementioned claim in Section III-B that our SAD framework is superior to GAN.

Effects of different components. Table VI summarizes the effects of different design components on Cityscapes [46] → Foggy Cityscapes [47]. The GPA [13] baseline is 39.5%. By adding the DC and SAR sequentially, we boost the mAP with an additional $+3.0\%$ and $+2.7\%$, achieving 42.5% and 45.2%,

(a) FoV$_x$= 40° *(b) FoV$_x$= 55°* *(c) FoV$_x$= 70°* *(d) FoV$_x$= 80°*

Fig. 6: Examples of diversified source images produced by our Domainness Creator with different FoV_xs.

TABLE VI: Ablation on Cityscapes to Foggy Cityscapes.

GPA [13]	DC	SAR	mAP	Gain
			39.5	-
√			42.5	3.0
√	√		45.2	5.7
√	√	√		

TABLE VII: Ablation of the domainnness-specific loss $\mathcal{L}_{spf}$.

Abaltions	mIoU
baseline(AdaptSegNet [26])	29.4
Ours (w/o $\mathcal{L}_{spf}$)	34.0
Ours (w $\mathcal{L}_{spf}$)	35.2

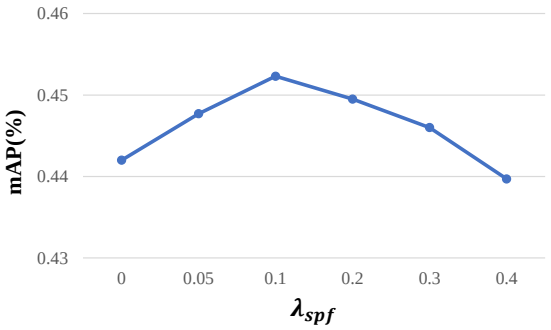


Fig. 7: Parameter analysis on the hyper-parameter λ_{spf}.

respectively. These improvements in object detection show the effects of individual components of our proposed approach. It also reveals that these two components are complementary and together they significantly promote the performance.

Effects of loss functions. Table VII shows the ablation of the domainnness-specific loss $\mathcal{L}_{spf}$ when adapting from Virtual KITTI to CKITTI for segmentation. The full framework with both DC and SAR can achieve 35.2 % mIoU. By removing the domainness-specific loss $\mathcal{L}_{spf}$ during the training process, the overall performance will drop by 1%. In addition, domainnness-invariant loss $\mathcal{L}_{inv}$ is critical for learning the domainnesss-invariant representations in the intra-domain adaptation, and cannot be removed. This shows that our SAR (the regularizer) needs to be trained under the guidance of both loss functions, *i.e.,* $\mathcal{L}_{spf}$ and $\mathcal{L}_{inv}$. Therefore, we cannot remove any of them.

F. Parameter Analysis

In this section, we investigate the sensitivity of the hyper-parameter λ_{spf} which balances the domain adaptation process. In Fig. 7, we plot the performance curve of models trained with different λ_{spf} values on the setting of Cityscapes → Foggy Cityscapes in object detection task. The highest mAP on target domain is achieved when the value of λ_{spf} is around 0.1, which means that this weight among different loss functions benefits domain adaptation the most. We simply set the same $\lambda_{spf} = 0.1$ in all experiments, to show the robustness of our method in different settings. Note that we use the original weighting ratio in [9], [10], [13], [16], [26], [29], [30] to balance $\mathcal{L}_{task}$ and $\mathcal{L}_{adv}$.

G. Qualitative Results

Fig. 4 visualizes the qualitative results of cross-domain object detection on two benchmarks, Cityscapes [46] → Foggy Cityscapes [47] and Cityscapes [46] → RTTS [48], respectively. As we can see from the pictures, the proposed method is able to produce high confidence predictions and is able to detect more objects when plugging into the current state-of-the-art methods, *e.g.,* GPA and SWDA [10].

Fig. 5 shows the qualitative results of cross-domain semantic segmentation from Virtual KITTI dataset [51] to CKITTI [46], [52]. With the aid of our proposed Self-Adversarial Disentangling framework, the models are able to produce correct

predictions at a high level of confidence, e.g., plugging it into AdaptSegNet [26]. As we can see from the figures, our method enables good performance on most categories, *e.g.,* 'vegetation', 'terrain', 'car', 'truck', and 'traffic sign' classes.

Fig. 6 shows the output results of Domainness Creator when receiving a source image at a time given the FoV gap. We get a series of diversified images with different $FoVx$s. From left to right it displays the processed image with FoV_x of $40°$, $55°$, $70°$ and $80°$, respectively. Due to the increased variations of domainness, a model trained on this domainness-diversified dataset is able to learn the domainness-invariant representation for feature alignment.

V. Conclusion

In this paper, we studied specific domain adaptation and proposed self-adversarial disentangling to learn the domainness-invariant feature in a specific dimension. The domainness creator aims to enrich the source domain and to provide additional supervisory signals for fully learning the domainness-invariant feature. The self-adversarial regularizer and two losses are introduced to narrow the intra-domain gap induced by different domainness. Extensive experiments validate our method on object detection and semantic segmentation under various domain-shift settings. Our method can be easily integrated into state-of-the-art architectures to attain considerable performance gains.

Acknowledgment

This work is supported by National Key Research and Development Program of China (No. 2019YFC1521104), National Natural Science Foundation of China (No. 61972157), Zhejiang Lab (No. 2020NB0AB01) and Shanghai Municipal Science, Technology Major Project (No. 2021SHZDZX0102) and Shanghai Science and Technology Commission (No. 21511101200). The author Qianyu Zhou is supported by Wu Wenjun Honorary Doctoral Scholarship, AI Institute, Shanghai Jiao Tong University.

References

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[2] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," vol. 28, 2015, pp. 91–99.

[4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[5] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2018.

[6] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[7] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[8] Z. Feng, Q. Zhou, Q. Gu, X. Tan, G. Cheng, X. Lu, J. Shi, and L. Ma, "Dmt: Dynamic mutual training for semi-supervised learning," *arXiv preprint arXiv:2004.08514*, 2020.

[9] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. 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*, 2018, pp. 3339–3348.

[10] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Strong-weak distribution alignment for adaptive object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6956–6965.

[11] Z. He and L. Zhang, "Multi-adversarial faster-rcnn for unrestricted object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6668–6677.

[12] X. Zhu, J. Pang, C. Yang, J. Shi, and D. Lin, "Adapting object detectors via selective cross-domain alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 687–696.

[13] M. Xu, H. Wang, B. Ni, Q. Tian, and W. Zhang, "Cross-domain detection via graph-induced prototype alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 355–12 364.

[14] C. Chen, Z. Zheng, X. Ding, Y. Huang, and Q. Dou, "Harmonizing transferability and discriminability for adapting object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8869–8878.

[15] C.-D. Xu, X.-R. Zhao, X. Jin, and X.-S. Wei, "Exploring categorical regularization for domain adaptive object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 724–11 733.

[16] Z. Shen, H. Maheshwari, W. Yao, and M. Savvides, "Scl: Towards accurate domain adaptive object detection via gradient detach based stacked complementary losses," *arXiv preprint arXiv:1911.02559*, 2019.

[17] C.-C. Hsu, Y.-H. Tsai, Y.-Y. Lin, and M.-H. Yang, "Every pixel matters: Center-aware feature alignment for domain adaptive object detector," in *European Conference on Computer Vision*. Springer, 2020, pp. 733–748.

[18] Z. He and L. Zhang, "Domain adaptive object detection via asymmetric tri-way faster-rcnn," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*. Springer, 2020, pp. 309–324.

[19] V. A. Sindagi, P. Oza, R. Yasarla, and V. M. Patel, "Prior-based domain adaptive object detection for hazy and rainy conditions," in *European Conference on Computer Vision*. Springer, 2020, pp. 763–780.

[20] G. Zhao, G. Li, R. Xu, and L. Lin, "Collaborative training between region proposal localization and classification for domain adaptive object detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 86–102.

[21] Y. Zheng, D. Huang, S. Liu, and Y. Wang, "Cross-domain object detection through coarse-to-fine feature adaptation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 766–13 775.

[22] Q. Gu, Q. Zhou, M. Xu, Z. Feng, G. Cheng, X. Lu, J. Shi, and L. Ma, "Pit: Position-invariant transform for cross-fov domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[23] W. Zhou, Y. Wang, J. Chu, J. Yang, X. Bai, and Y. Xu, "Affinity space adaptation for semantic segmentation across domains," *IEEE Transactions on Image Processing*, vol. 30, pp. 2549–2561, 2020.

[24] Y. Luo, P. Liu, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Category-level adversarial adaptation for semantic segmentation using purified features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.

[25] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *International conference on machine learning*. PMLR, 2018, pp. 1989–1998.

[26] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7472–7481.

[27] Y. Li, L. Yuan, and N. Vasconcelos, "Bidirectional learning for domain adaptation of semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6936–6945.

[28] M. Kim and H. Byun, "Learning texture invariant representation for domain adaptation of semantic segmentation," in *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12 975–12 984.

[29] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2507–2516.

[30] Z. Wang, M. Yu, Y. Wei, R. Feris, J. Xiong, W.-m. Hwu, T. S. Huang, and H. Shi, "Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 635–12 644.

[31] W.-L. Chang, H.-P. Wang, W.-H. Peng, and W.-C. Chiu, "All about structure: Adapting structural information across domains for boosting semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1900–1909.

[32] Y. Yang and S. Soatto, "Fda: Fourier domain adaptation for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4085–4095.

[33] Z. Lu, Y. Yang, X. Zhu, C. Liu, Y.-Z. Song, and T. Xiang, "Stochastic classifiers for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9111–9120.

[34] Y. Yang, D. Lao, G. Sundaramoorthi, and S. Soatto, "Phase consistent ecological domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9011–9020.

[35] F. Pan, I. Shin, F. Rameau, S. Lee, and I. S. Kweon, "Unsupervised intra-domain adaptation for semantic segmentation through self-supervision," in *Unsupervised Intra-domain Adaptation for Semantic Segmentation through Self-Supervision*, 2020, pp. 3764–3773.

[36] J. Yang, R. Xu, R. Li, X. Qi, X. Shen, G. Li, and L. Lin, "An adversarial perturbation oriented domain adaptation approach for semantic segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 613–12 620.

[37] H. Wang, T. Shen, W. Zhang, L. Duan, and T. Mei, "Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation," in *European conference on computer vision*, vol. 12359. Springer, 2020, pp. 642–659.

[38] T. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "DADA: depth-aware domain adaptation in semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7363–7372.

[39] J. Choi, T. Kim, and C. Kim, "Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6830–6840.

[40] Y. Luo, P. Liu, T. Guan, J. Yu, and Y. Yang, "Significance-aware information bottleneck for domain adaptive semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6778–6787.

[41] Y. Zou, Z. Yu, X. Liu, B. Kumar, and J. Wang, "Confidence regularized self-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5982–5991.

[42] Y. Zou, Z. Yu, B. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 289–305.

[43] X. Zhu, H. Zhou, C. Yang, J. Shi, and D. Lin, "Penalizing top performers: Conservative loss for semantic segmentation adaptation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 568–583.

[44] Y.-H. Tsai, K. Sohn, S. Schulter, and M. Chandraker, "Domain adaptation for structured output via discriminative patch representations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[45] S. Guo, Q. Zhou, Y. Zhou, Q. Gu, J. Tang, Z. Feng, and L. Ma, "Label-free regional consistency for image-to-image translation," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2021, pp. 1–6.

[46] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. CVPR*, 2016, pp. 3213–3223.

[47] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *International Journal of Computer Vision*, vol. 126, no. 9, pp. 973–992, 2018.

[48] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang, "Benchmarking single-image dehazing and beyond," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 492–505, 2018.

[49] C. Sakaridis, D. Dai, S. Hecker, and L. Van Gool, "Model adaptation with synthetic and real data for semantic dense foggy scene understanding," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 687–704.

[50] X. Hu, C.-W. Fu, L. Zhu, and P.-A. Heng, "Depth-attentional features for single-image rain removal," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8022–8031.

[51] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4340–4349.

[52] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *IJR*, vol. 32, no. 11, pp. 1231–1237, 2013.

[53] Q. Zhou, Z. Feng, Q. Gu, J. Pang, G. Cheng, X. Lu, J. Shi, and L. Ma, "Context-aware mixup for domain adaptive semantic segmentation," *arXiv preprint arXiv:2108.03557*, 2021.

[54] Q. Zhou, Z. Feng, Q. Gu, G. Cheng, X. Lu, J. Shi, and L. Ma, "Uncertainty-aware consistency regularization for cross-domain semantic segmentation," *arXiv preprint arXiv:2004.08878*, 2020.

[55] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2517–2526.

[56] Z. Feng, Q. Zhou, G. Cheng, X. Tan, J. Shi, and L. Ma, "Semi-supervised semantic segmentation via dynamic self-training and classbalanced curriculum," *arXiv preprint arXiv:2004.08514*, vol. 1, no. 2, p. 5, 2020.

[57] G. French, M. Mackiewicz, and M. Fisher, "Self-ensembling for visual domain adaptation," in *Proceedings of the International Conference on Learning Representations*, 2018.

[58] C. S. Perone, P. Ballester, R. C. Barros, and J. Cohen-Adad, "Unsupervised domain adaptation for medical imaging segmentation with self-ensembling," *NeuroImage*, vol. 194, pp. 1–11, 2019.

[59] V. Olsson, W. Tranheden, J. Pinto, and L. Svensson, "Classmix: Segmentation-based data augmentation for semi-supervised learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1369–1378.

[60] W. Tranheden, V. Olsson, J. Pinto, and L. Svensson, "Dacs: Domain adaptation via cross-domain mixed sampling," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1379–1389.

[61] L. Melas-Kyriazi and A. K. Manrai, "Pixmatch: Unsupervised domain adaptation via pixelwise consistency training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 435–12 445.

[62] N. Araslanov and S. Roth, "Self-supervised augmentation consistency for adapting semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 384–15 394.

[63] Y. Xu, B. Du, L. Zhang, Q. Zhang, G. Wang, and L. Zhang, "Self-ensembling attention networks: Addressing domain shift for semantic segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 5581–5588.

[64] T. Kim, M. Jeong, S. Kim, S. Choi, and C. Kim, "Diversify and match: A domain adaptive representation learning paradigm for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 456–12 465.

[65] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," vol. 27, 2014.

[66] X. Yue, Y. Zhang, S. Zhao, A. Sangiovanni-Vincentelli, K. Keutzer, and B. Gong, "Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2100–2110.

[67] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European conference on computer vision*, 2018, pp. 172–189.

[68] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 35–51.

[69] K. Ridgeway and M. C. Mozer, "Learning deep disentangled embeddings with the f-statistic loss," 2018.

[70] T. R. Scott, K. Ridgeway, and M. C. Mozer, "Adapted deep embeddings: A synthesis of methods for k-shot inductive transfer learning," 2018.

[71] Y.-C. Liu, Y.-Y. Yeh, T.-C. Fu, S.-D. Wang, W.-C. Chiu, and Y.-C. F. Wang, "Detach and adapt: Learning cross-domain disentangled deep representation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8867–8876.

[72] H. Wang, H. He, and D. Katabi, "Continuously indexed domain adaptation," in *The International Conference on Machine Learning*, 2020.

[73] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[74] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*, 2015, pp. 1180–1189.

[75] H. Zhang and V. M. Patel, "Densely connected pyramid dehazing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3194–3203.

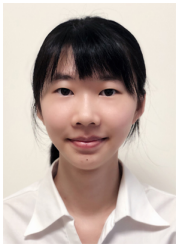
[76] X. Liu, Y. Ma, Z. Shi, and J. Chen, "Griddehazenet: Attention-based multi-scale network for image dehazing," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7314–7323.

[77] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[78] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[79] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

Qianyu Zhou is currently pursuing his Ph.D. degree in the Department of Computer Science and Engineering, Shanghai Jiao Tong University. Before that, he received a B.Sc. degree in Jilin University in 2019. His current research interests focus on computer vision, scene understanding, domain adaptation.



Qiqi Gu received a B.E. degree in Shanghai Jiao Tong University in 2015 and is now a second-year master student in Department of Computer Science and Engineering, Shanghai Jiao Tong University. Her current research interests focus on domain adaptation of object detection and semantic segmentation.

Jiangmiao Pang is currently a Postdoctoral Research Fellow at Multimedia Laboratory, the Chinese University of Hong Kong. He obtained his Ph.D. degree from Zhejiang University in 2021. His research interests include computer vision and robotics, especially their applications in autonomous driving.

Zhengyang Feng is currently pursuing his M.Sc. degree in the Department of Computer Science and Engineering, Shanghai Jiao Tong University. Before that, he received a B.E. degree in information security from Harbin Institute of Technology, Weihai, China, in 2020. His current research interests focus on pattern recognition with limited human supervision.

Guangliang Cheng is currently a Senior Research Manager in SenseTime. Before that, he was a Postdoc researcher in the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, China, and he received his Ph.D. degree with national laboratory of pattern recognition (NLPR) from the Institute of Automation, Chinese Academy of Sciences, Beijing. His research interests include autonomous driving, scene understanding, domain adaptation and remote sensing image processing.

Xuequan Lu is an Assistant Professor at the School of Information Technology, Deakin University, Australia. He spent more than two years as a Research Fellow in Singapore. Prior to that, he earned his Ph.D at Zhejiang University (China) in June 2016. His research interests mainly fall into the category of visual computing, for example, geometry modeling, processing and analysis, animation/simulation, 2D data processing and analysis. More information can be found at http://www.xuequanlu.com.

Jianping Shi is an Executive Research Director at SenseTime. Currently her team works on developing algorithms for autonomous driving, scene understanding, remote sensing, etc. She got her Ph.D. degree in Computer Science and Engineering Department in the Chinese University of Hong Kong in 2015 under the supervision of Prof. Jiaya Jia. Before that, she received the B. Eng degree from Zhejiang University in 2011. She has served regularly on the organization committees of numerous conferences, such as Area Chair of CVPR 2020, ICCV 2019, etc.

Lizhuang Ma received his B.S. and Ph.D. degrees from the Zhejiang University, China in 1985 and 1991, respectively. He is now a Distinguished Professor, Ph.D. Tutor, and the Head of the Digital Media and Computer Vision Laboratory at the Department of Computer Science and Engineering, Shanghai Jiao Tong University, China. He was a Visiting Professor at the Frounhofer IGD, Darmstadt, Germany in 1998, and was a Visiting Professor at the Center for Advanced Media Technology, Nanyang Technological University, Singapore from 1999 to 2000. He has published more than 200 academic research papers in both domestic and international journals. His research interests include computer aided geometric design, computer graphics, computer vision, scientific data visualization, computer animation, digital media technology, and theory and applications for computer graphics, CAD/CAM. He serves as the reviewer of IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Image Processing, IEEE Transactions on Multimedia, etc.