SiamDoGe: Domain Generalizable Semantic Segmentation using Siamese Network

Zhenyao Wu[1], Xinyi Wu[1], Xiaoping Zhang[2], Lili Ju[1,†], and Song Wang[1,†]

[1] University of South Carolina
{zhenyao, xinyiw}@email.sc.edu ju@math.sc.edu songwang@cec.sc.edu
[2] Wuhan University
xpzhang.math@whu.edu.cn

Abstract. Deep learning-based approaches usually suffer from performance drop on out-of-distribution samples, therefore domain generalization is often introduced to improve the robustness of deep models. Domain randomization (DR) is a common strategy to improve the generalization capability of semantic segmentation networks, however, existing DR-based algorithms require collecting auxiliary domain images to stylize the training samples. In this paper, we propose a novel domain generalizable semantic segmentation method, "**SiamDoGe**", which builds upon a DR approach without using auxiliary domains and employs a Siamese architecture to learn domain-agnostic features from the training dataset. Particularly, the proposed method takes two augmented versions of each training sample as input and produces the corresponding predictions in parallel. Throughout this process, the features from each branch are randomized by those from the other to enhance the feature diversity of training samples. Then the predictions produced from the two branches are enforced to be consistent conditioned on feature sensitivity. Extensive experiment results demonstrate the proposed method exhibits better generalization ability than existing state-of-the-arts across various unseen target domains.

Keywords: Domain generalization; Semantic segmentation; Siamese Network; Domain randomization

1 Introduction

Semantic image segmentation associates each pixel to a semantic label and has a wide range of applications in real world, such as autonomous driving [65,5], robotic navigation [53,41] and medical image diagnostic [40,68]. Current deep learning-based approaches [34,65,5] have achieved very promising results through training on large-scale labeled datasets [13,11,1], but these datasets are usually very laborious to collect and annotate. Another well-known phenomenon is that a deep model trained on one dataset often fits well on its own test split (in-domain) but suffers from a huge performance drop on other datasets (cross-domain), and this phenomenon is usually called *domain shift*.

† Co-corresponding authors. Code is available at github.com/W-zx-Y/SiamDoGe.

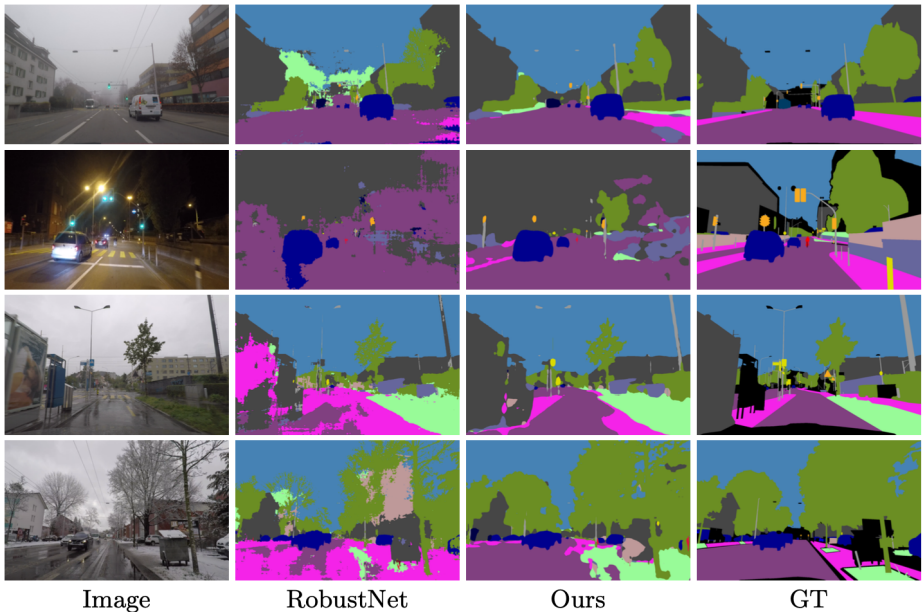

Fig. 1. Some visual comparisons of the domain generalizable semantic segmentation results produced by the proposed method and RobustNet [10]. Both models are trained on the GTAV dataset and evaluated using the ACDC dataset with four conditions of Fog, Night, Rain and Snow, as shown in rows 1 to 4 respectively.

In recent years, domain adaptive (DA) semantic segmentation [22,8,55,57,44,60,63] was proposed to bridge domain gaps so that a model trained on one domain (source) also works well on other domains (target). This is achieved by leveraging multiple unlabeled target samples as references while training the source with supervision. However, this approach has two limitations: 1) target samples are always required even if it can be as few as one [36]; and 2) multiple times of adaptation are required to be performed when there are multiple desired target domains. Compared with DA, domain generalization (DG) is a more universal solution to handle arbitrary domain shifts thus does not have a preference towards a particular target domain. It aims to reduce the model sensitivity to the change of data distribution via domain-agnostic feature learning. DG has been typically studied with two different settings: multi-source DG [42,14,29,31,30,67] and single-source DG [46,66,58]. In this paper, we study single-source DG for semantic segmentation.

Existing domain generalizable semantic segmentation approaches are mainly based on feature normalization [45,10] and domain randomization [62,23,27] (DR). It is observed that the domain randomization-based approaches usually can achieve better generalization capacity than the domain normalization-based ones due to the use of auxiliary real-world domains, *e.g.*, ImageNet [12] or web-

crawled images, for source image stylization. On the other hand, the DR-based methods also have the following drawbacks: 1) their DG performances highly depend on the choice of auxiliary domains and it takes a lot of time to carefully collect data in the domains related to the task in order to avoid impure DG [27]; and 2) most of them lack enough control and could undesirably alter the semantic structures (domain invariant features) of images [23].

With this observation, we propose in this paper a novel domain generalizable semantic segmentation method, "SiamDoGe", which is based on domain randomization but does not use other auxiliary domains. Our work is partially inspired by SimSiam [7], a Siamese network for unsupervised representation learning by comparing two views of one image. In the proposed method, two augmented versions of a source sample are first generated, then a Siamese network is employed to find the crucial shared invariant representations from the two branches for domain generalization. Specifically, the features from the two branches are randomized interdependently during training. There are two natural advantages for such design over existing DR-based algorithms: 1) collecting extra data in auxiliary domains is no longer needed; and 2) more controllability is obtained since the randomization is performed by using two images that share common content. Besides, we also study the feature sensitivity by comparing low-level features from the two branches. The prediction consistency of the two branches is then enforced with more attention being paid to more sensitive regions since it is usually difficult to obtain domain-agnostic features in those regions. Extensive experimental results verify the effectiveness of our approach and show that the proposed SiamDoGe generalizes very well to multiple unseen domains both qualitatively (see Figure 1 for sample visual comparisons) and quantitatively.

The main contributions of our work are summarized as follows:

- A new domain generalizable semantic segmentation approach, "SiamDoGe", is developed for domain-invariant feature learning by employing Siamese structure.
- Our method achieves a more controllable self-guided randomization without using auxiliary domains, and better domain-agnostic features by taking account of feature sensitivity when ensuring the prediction consistency.
- We evaluate the performance on various unseen domains and the results show that our method exhibits better generalization capacity than existing state-of-the-arts.

2 Related works

In this section, we discuss the related works on DA- and DG-based semantic segmentation and the background of Siamese networks.

Domain adaptive semantic segmentation Domain adaptation (DA) is related to our work since its goal is to minimize domain discrepancy. There are

The auxiliary domains should not share common data samples with the unseen target domain according to the definition of DG.

two commonly-used strategies for domain adaptive semantic segmentation: adversarial training [22,55,21,35,57,44] and self-training [69,70,33,60,38,26,63,17]. The former usually trains a discriminator and a segmentation network alternatively to align source and target domains and the latter leverages the confident pseudo labels to achieve more performance gains via multiple rounds of retraining on the target domain. A lot of DA scenarios have been studied for semantic segmentation such as synthetic-to-real [55,57,69,70,60,44,63], cross-time of day [49,50,51,59], cross weathers [51], cross cities [22,55], and many of them benefit autonomous driving. However, retraining is required whenever a new DA scenario (target) appears. Differently, in this paper, we explore domain generalization, which is a more universal solution than DA for handling the domain discrepancy since it does not require to specify a target domain.

Domain generalizable semantic segmentation Existing DG semantic segmentation is usually achieved by specific designs of feature normalization [45,10], knowledge distillation [6], or domain randomization [62,23,27].

IBN-Net is the first DG semantic segmentation approach [45] integrating instance normalization [56] and batch normalization [25], where the former learns appearance-invariant features and the latter preserves content information. This work was recently extended by incorporating an instance selective whitening loss to selectively remove the feature co-variances that respond sensitively to the domain shift in [10]. In [6], Chen *et al.*formulate the synthetic-to-real generalization as a lifelong learning problem by enforcing the representation similarity between synthetically trained models and the ImageNet pre-trained model via a distillation loss.

Domain randomization (DR) is a more frequently used strategy for DG. Yue *et al.*[62] first explored the DR for semantic segmentation where auxiliary-domain images are carefully picked from ImageNet [12] to stylize the training images and the prediction consistency is enforced across all stylized images of one training sample. However, DR is not controllable and might hurt the domain invariant features [23] crucial to DG. Huang *et al.*[23] further refined the DR strategy by transferring images into the frequency domain and only randomizing the domain-variant frequency components with the domain-invariant ones unchanged. Very recently, Kim *et al.*[27] proposed a non-parametric style injection module to randomize the training images on-the-fly using a large amount of web-crawled images which are real and related to the application of autonomous driving. In general, our proposed method falls in the group of DR but does not use auxiliary domains for source image stylization and thus is more controllable.

Siamese network Siamese neural networks [4] were proposed to learn semantic similarity and have been shown to work well on various vision tasks such as object-tracking [3,54,18,64,9], image co-segmentation [32,2], one-shot learning [28] and unsupervised visual representation learning [19,16,7], *etc.*. By definition, Siamese networks are weight-sharing neural networks applied to two or more inputs for comparing the entities [4,7]. It has been employed by many of existing DA approaches [22,55,57,39] to bridge the domain gap between source and target

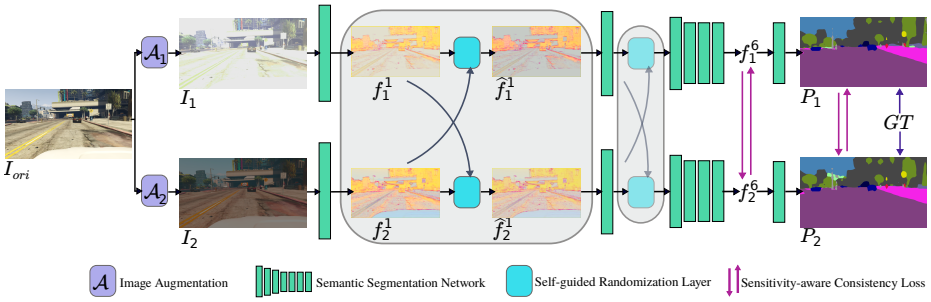

Fig. 2. The overall architecture and training pipeline of the proposed SiamDoGe. A training sample I_{ori} is first taken as input to obtain two augmented versions I_1 and I_2, respectively. Then each branch i, $(i = 1, 2)$ is fed with an augmented image I_i and produces a prediction P_i with a group of intermediate features $\{f_i^j\}_{j=1}^6$. The proposed self-guided randomization is particularly applied to the features obtained from the first and second layers of the segmentation network, *i.e.*, f_i^1 and f_i^2, to produce the corresponding randomized features $\widehat{f}_i^1$ and $\widehat{f}_i^2$. The semantic segmentation network shares weights across the two branches and the whole pipeline is trained via the standard cross-entropy loss and regularized by a novel sensitivity-guided consistency loss.

images. In this paper, we propose a novel DG semantic segmentation approach based on the Siamese architecture.

3 Proposed Method

3.1 Overview

Our model improves the performance and robustness of semantic segmentation networks via two specially designed components: a *self-guided randomization* and a *sensitivity-guided consistency training*. The former allows to perform domain randomization with the training sample itself, and the latter encourages to find domain-invariant features based on feature sensitivity by comparing the two branches.

The overall architecture and training pipeline of the proposed SiamDoGe is illustrated in Figure 2. Given a training sample image I_{ori}, we first obtain two augmented views I_1 and I_2 for it with two random simple color jittering transformations $\mathcal{A}_1$ and $\mathcal{A}_2$, respectively. The two views which share the same content but different visual styles are then fed into a weight-sharing Siamese network (two branches) for semantic segmentation. During feature extraction, the proposed self-guided randomization is particularly applied to the features obtained from the first and second layers of the segmentation network, *i.e.*, f_i^1 and f_i^2, to produce the corresponding randomized features $\widehat{f}_i^1$ and $\widehat{f}_i^2$ for each branch $i \in \{1, 2\}$. The whole pipeline is trained under the supervision of source domain ground truths for both branches via the standard cross-entropy loss plus a novel sensitivity-guided consistency loss.

3.2 The self-guided randomization

To achieve more controllable domain randomization without accessing to auxiliary domains, we propose to randomize the source image in a self-guided way, which is implemented via feature normalization inspired by [67]. The details of this operation are shown in Figure 3.

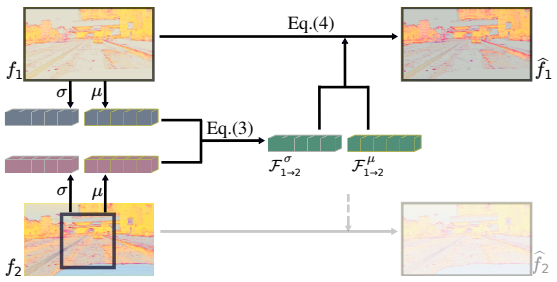

Fig. 3. An illustration of the proposed self-guided randomization process.

Specifically, the inputs of the self-guided randomization layer are two intermediate feature maps f_1 and $f_2 \in \mathbb{R}^{C\times H\times W}$ from the two branches with C, H and W representing channel, height and width, respectively. Following [24], we first denote the spatial feature statistics $\mu(.)$ and $\sigma(.) \in \mathbb{R}^C$ of a feature f by

$$\mu(f)_{(c)} = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} f_{(c,h,w)}, \tag{1}$$

and

$$\sigma(f)_{(c)} = \sqrt{\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}(f_{(c,h,w)} - \mu(f)_{(c)})^2 + \epsilon}, \tag{2}$$

where ϵ is set to 10^{-10}.

A naive randomization can be easily obtained via adding Gaussian noise to the source feature statistics, however, such perturbation is also lacking in control and still might destroy the domain-invariant features. Based on the concept of "domain flow" introduced in [15] to describe intermediate domains between the source and target domains for domain adaptation, we propose a concept of "intra-domain flow" which represents intermediate intra-domains within the source domain for domain generalization. We first define the intra-domain flow from f_1 to f_2 based on their feature statistics as:

$$\mathcal{F}^{\mu}_{1\to 2} = \mu(\mathcal{C}(f_2)) - \mu(f_1); \quad \mathcal{F}^{\sigma}_{1\to 2} = \sigma(\mathcal{C}(f_2)) - \sigma(f_1), \tag{3}$$

where the function $\mathcal{C}$ stands for the random cropping operation with a size of 64×64 used to help improve the diversity of the flow. Then the computed intra-

domain flow $\{\mathcal{F}^{\mu}_{1\to2}, \mathcal{F}^{\sigma}_{1\to2}\}$ is adopted to randomize f_1 as follows:

$$\begin{aligned}\hat{f}_1 &= (\sigma(f_1) + \lambda\mathcal{F}^{\sigma}_{1\to2})\left(\frac{f_1 - \mu(f_1)}{\sigma(f_1)}\right) \\ &\quad +(\mu(f_1) + \lambda\mathcal{F}^{\mu}_{1\to2}) \\ &= f_1 + \lambda\left(\mathcal{F}^{\sigma}_{1\to2}\frac{f_1 - \mu(f_1)}{\sigma(f_1)} + \mathcal{F}^{\mu}_{1\to2}\right),\end{aligned} \tag{4}$$

where $\lambda \in [0, 1]$ is a randomly generated hyper-parameter used to control the randomization. Similarly, we also compute the flow $\{\mathcal{F}^{\mu}_{2\to1}, \mathcal{F}^{\sigma}_{2\to1}\}$ and then randomize f_2 via Eq. (3) and Eq. (4) by switching f_1 and f_2, and replacing $1 \to 2$ with $2 \to 1$.

3.3 The sensitivity-guided consistency training

Consistency training was explored in [62] for domain generalizable semantic segmentation, where various stylized source images of one training sample in auxiliary domains are enforced to have consistent predictions. In this paper, we make one further step to propose a sensitivity-guided consistency training to ensure the consistency of the two branches as illustrated in Figure 4. Our insight is that the difference between low-level features from the two augmented versions can well describe how the features are sensitive to the "domain" shift, *i.e.*, a small difference means the feature is robust while a large one means it is variant to the shift. Therefore, the proposed loss function will pay more attention to the sensitive regions and less attention to the insensitive regions that are already generalized well.

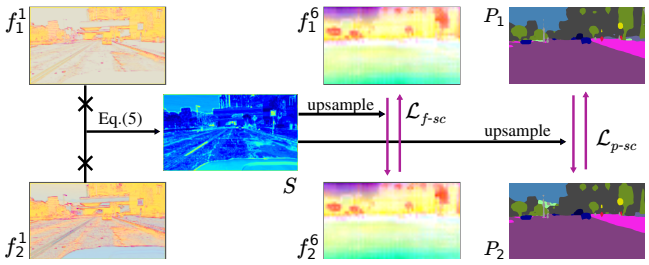

Fig. 4. An illustration of the proposed sensitivity-guided consistency training. In a sensitivity map S, the darker blue regions are less sensitive than the light blue ones.

We first measure the distance between the low-level features f_1^1 and f_2^1 from the two branches and obtain the feature sensitivity map $S \in \mathbb{R}^{H\times W}$ by

$$S_{(h,w)} = \zeta\left(\frac{1}{C}\sum_{c=1}^{C}|f^1_{1,(c,h,w)} - f^1_{2,(c,h,w)}|\right), \tag{5}$$

where ζ is the stop-gradient operation that is borrowed from [7]. S then is involved in consistency loss computation. Specifically, we build the sensitivity-guided consistency loss in both the feature level ($\mathcal{L}_{f-sc}$) and the prediction level ($\mathcal{L}_{p-sc}$), which are formulated as:

$$\mathcal{L}_{f-sc} = \frac{1}{CHW} \sum_{h,w} \left(S_{(h,w)} \sum_{c=1}^{C} |f^6_{1,(c,h,w)} - f^6_{2,(c,h,w)}| \right), \tag{6}$$

$$\mathcal{L}_{p-sc} = \frac{1}{CHW} \sum_{h,w} \left(S_{(h,w)} \sum_{c=1}^{C} |P_{1,(c,h,w)} - P_{2,(c,h,w)}| \right), \tag{7}$$

where the feature f^6_i and the prediction P_i are upsampled to the same resolution as S.

3.4 Implementation details

Architecture Following [10], we use the DeeplabV3+ [5] as our semantic segmentation network with Resnet-50 [20] as the backbone of the proposed SiamDoGe. It contains seven layers in total with the first one being a combination of Conv-BN-ReLU-MAX POOLING operations, the 2-5th ones being the residual blocks, the 6th one being the ASPP layer with the output stride of 16 and the last one is the classifier layer.
Loss function The training objective of SiamDoGe is a combination of semantic segmentation loss and the sensitivity-guided consistency loss defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \alpha(\mathcal{L}_{f-sc} + \mathcal{L}_{p-sc}), \tag{8}$$

where α is a weighting parameter for balancing the two loss terms.
Augmentation Our SiamDoGe only uses the random color jittering for input image augmentation, *i.e.*, $\mathcal{A}_1$ and $\mathcal{A}_2$ in Figure 2, which produces two different views of the same image. Following [10], the parameters for color jittering are set to 0.8, 0.8, 0.8 and 0.3 for brightness, contrast, saturation and hue, respectively.
Optimization The SGD optimizer is used with an initial learning rate of 10^{-2} and a momentum of 0.9. The polynomial learning rate scheduler [65] with the power of 0.9 is also applied to stabilize the training. The whole network is trained using two Nvidia Tesla V100 GPU cards with a batch size of 8 for 40K iterations in total for all experiments and each experiment requires around 22 hours for training. The hyper-parameter α is set to 0 for the first 10K iterations and 10 for the rest iterations during training.

4 Experiments

In this section, we first introduce all the datasets that are involved in the experiments. Then we demonstrate the excellent performance of our SiamDoGe by comparing it with existing state-of-the-arts, and also empirically study the effects of its key components via ablation studies.

4.1 Datasets

We evaluate the proposed SiamDoGe on two DG settings based on different domains, including GTAV(G) → {Cityscapes(C), SYNTHIA(S), Mapillary(M), BDD-100K(B), ACDC(A)} and Cityscapes(C) → {GTAV(G), SYNTHIA(S), Mapillary(M), BDD-100K(B), ACDC(A)}. We adopt the mean intersection-over-union (mIoU) as the evaluation metric (the higher the better) and all the datasets are evaluated based on 19 classes defined by Cityscapes.

GTAV [47] is a large-scale self-annotated synthetic dataset collected from commercial video games. It contains images with a resolution of 1,914×1,052 and is divided into 12,403/6,382/6,181 for training, validation and testing purposes. Here we use the training and validation images with their labels for training when it serves as the seen domain and only use its validation set when it is treated as an unseen domain.

Cityscapes [11] captures real-world urban street scenes from different cities and it provides 5,000 high quality manually-annotated images in pixel level with a resolution of 2,048×1,024. The images are split into subsets of 2,975/500/1,525 images for training, validation and testing, respectively. Here we use the training set when it serves as the seen domain and the validation set when it serves as an unseen domain.

SYNTHIA [48] is a synthetic dataset consisting of 9,400 self-labeled images with a resolution of 960×720. Here we use the 2,820 images split by [10] for evaluation.

Mapillary [43] is a real-world dataset that is designed to capture scenes in the wild variations across season/weather conditions, viewing perspectives, time and resolution (at least 1,920×1,080), *etc.* Here we use its original validation split (2,000 images) for evaluation.

BDD-100K [61] is an another real-world dataset recorded in diverse weather conditions at different times of the day. The resolution of the images is 1,280×720. Here we use its original validation split (1,000 images) for evaluation.

ACDC [51] is the largest adverse condition dataset for semantic segmentation to date. Different from Mapillary and BDD-100K which also contain normal condition scenes, ACDC purely consists of images with four common adverse conditions of fog, nighttime, rain and snow. The resolution of the images is 1,920×1,080. Here we use its validation set (406 images) for evaluation, *i.e.,* 100/100/100/106 for fog/rain/snow/nighttime conditions, respectively.

4.2 Comparison with state-of-the-arts

We first compare the performance of our SiamDoGe with several existing DG approaches for semantic segmentation under the setting G → {C, B, M, S, A}, including the feature normalization-based ones (not using auxiliary domains): IBN-Net [45] and RobustNet [10], and the domain randomization-based ones (using auxiliary domains): DRPC [62] and WEDGE [27]. Table 1 reports all the quantitative comparison results. Note that only the results on G → {C, B, M}

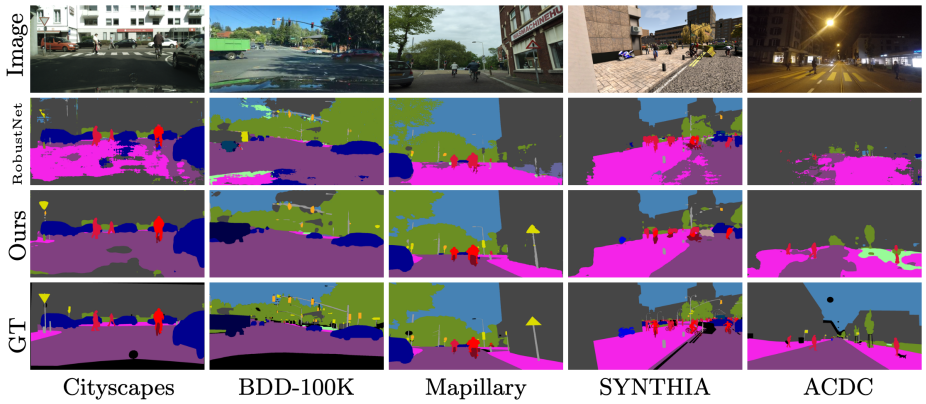

Fig. 5. Qualitative results of SiamDoGe and RobustNet [10] under the setting G → {C, B, M, S, A}.

Table 1. Quantitative comparison results of our SiamDoGe and the existing state-of-the-art DG approaches for semantic segmentation under the setting G → {C, B, M, S, A}. All the methods use ResNet-50 as backbone. The *avg.* represents the average mIoU (%) over the five datasets. †indicates that auxiliary domains are required. The best results are highlighted with **bold**.

Methods	C	B	M	S	A	*avg.*
DRPC †[62]	37.41	32.14	34.12	-	-	-
WEDGE †[27]	38.15	36.14	**43.21**	-	-	-
IBN-Net [45]	33.85	32.30	37.75	27.90	22.55	30.87
RobustNet [10]	36.58	35.20	40.33	28.30	25.46	33.14
SiamDoGe (ours)	**42.96**	**37.54**	40.64	**28.34**	**29.25**	**35.75**

are available for DRPC and WEDGE in the literature. It can be observed that our SiamDoGe achieves the best performance across all unseen target domains among the methods of not using auxiliary domains and surpasses the second best by a large margin on Cityscapes (6.38 mIoU) and ACDC (3.79 mIoU). More surprisingly, our SiamDoGe also outperforms the methods of using auxiliary domains, *e.g.*, DRPC significantly on all the three unseen target domains and WEDGE slightly on Cityscapes and BDD-100K. It is worth mentioning that WEDGE uses around 5K auxiliary domain images for randomization which is even larger than those in the unseen target domains. Some qualitative comparison results with RobustNet under this setting are provided in Figure 5, where we observe that our method achieves better results visually for the truck in the sample from BDD-100K, the traffic sign in the sample from Mapillary and the person in the sample from ACDC.

Table 2. Quantitative comparison results of SiamDoGe and the existing state-of-the-art DG approaches without using auxiliary domains for semantic segmentation under the setting C → {G, B, M, S, A}. All the methods use ResNet-50 as backbone.

Methods	G	B	M	S	A	*avg.*
IBN-Net [45]	45.06	48.56	57.04	26.14	44.05	44.17
RobustNet [10]	45.00	50.73	58.64	26.20	46.91	45.50
SiamDoGe	**45.08**	**51.53**	**59.00**	**26.67**	**52.34**	**46.92**

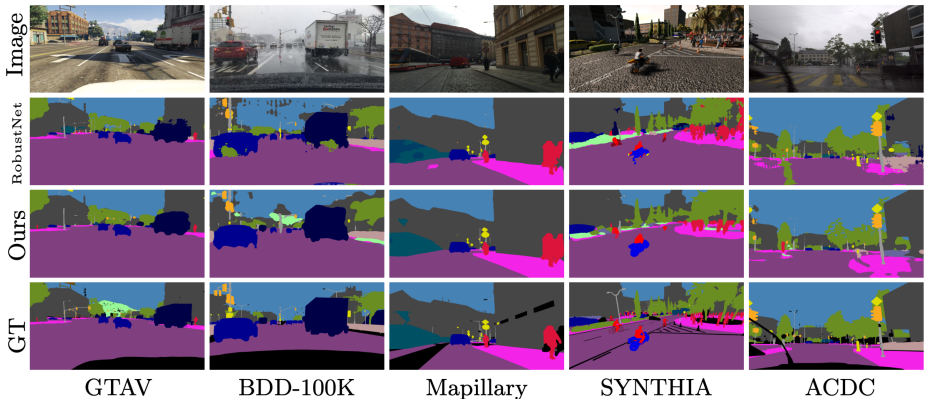

Fig. 6. Qualitative comparison results of SiamDoGe and RobustNet [10] under the setting C → {G, B, M, S, A}.

Similarly, we compare the performance of our SiamDoGe with the state-of-the-arts [45,10] under the setting C → {G, B, M, S, A}. The quantitative results are reported in Table 2 with some visualization results shown in Figure 6. Consistently, our SiamDoGe wins across all unseen domains again and surpasses the second best [10] on ACDC dataset by 5.44 mIoU. In addition, from Figure 6 we find that our method obtains better predictions especially for the classes of track, car, motor, train and *etc.*.

For a fair comparison, we also report the comparison of computational costs of the proposed SiamDoGe with RobustNet [10] in Table 3. It is observed that our method is more efficient than RobustNet.

Table 3. Comparison of computational costs with ResNet-50 as backbone.

Methods	# of Params	GFLOPs	Inference Time (ms)
RobustNet	40.35	60.69	7.58
SiamDoGe	40.23	43.00	5.71

Table 4. Quantitative comparison results of our SiamDoGe and the existing state-of-the-art DG approaches without using auxiliary domains for semantic segmentation under the setting G → {C, B, M, S, A}. All the methods use ShuffleNetV2 as backbone.

Methods	C	B	M	S	A	*avg.*
IBN-Net [45]	27.10	31.82	34.89	**25.56**	22.33	28.34
RobustNet [10]	30.98	32.06	35.31	24.31	21.27	28.79
SiamDoGe	**34.40**	**34.23**	**35.87**	21.95	**25.22**	**30.33**

Table 5. Quantitative comparison results of our SiamDoGe and the existing state-of-the-art DG approaches without using auxiliary domains for semantic segmentation under the setting G → {C, B, M, S, A}. All the methods use MobileNetV2 as backbone.

Methods	C	B	M	S	A	*avg.*
IBN-Net [45]	30.14	27.66	27.07	**24.98**	20.30	26.03
RobustNet [10]	30.86	30.05	30.67	24.43	23.26	27.85
SiamDoGe	**34.15**	**34.50**	**32.34**	23.53	**24.17**	**29.74**

Other backbones Following [10], we also employ ShuffleNetV2 [37] and MobileNetV2 [52] as additional backbones for performance evaluation. The models are compared under the setting G → {C, B, M, S, A} with corresponding results reported in Table 4 and Table 5, respectively. We observe that our method still achieves the best performance on average over all the unseen domains. Among all five test domains, our SiamDoGe achieves the best on the four real-world domains for both backbones.

4.3 Ablation studies

On main model components We first examine how each of our model components impacts the DG performance for semantic segmentation by testing several model variants. The numerical results obtained under both the settings G → {C, B, M, S, A} and C → {G, B, M, S, A} are shown in Table 6. "Single branch" serves as the baseline of our method by feeding the randomly augmented images (using $\mathcal{A}$) into only one branch to produce the segmentation results. "Siamese Network" means two branches take two augmented views of a sample as input and the predictions from the two branches are supervised by the ground-truth independently. The third one further models the relationship between the branches with the proposed sensitivity-guided consistency. The last one is our full model by adding the self-guided randomization. By comparing the former three variants, we can find that the Siamese structure is meaningless without modeling the consistency, *i.e.*, just doubles the batch size of the "Single branch". For the two settings, the sensitivity-guided consistency brings 2.12/1.17 mIoU gains on average and the self-guided randomization further brings 1.50/1.75 mIoU gains.

Table 6. Ablation study for main components of our SiamDoGe, including the Siamese network, the sensitivity-guided consistency (SC) and the self-guided randomization (SR).

	Trained on Cityscapes						Trained on GTAV					
Variants	G	B	M	S	A	*avg.*	C	B	M	S	A	*avg.*
Single branch	42.27	47.02	55.05	24.66	45.78	42.96	40.75	33.69	37.09	29.42	23.67	32.92
+ Siamese Network	40.98	46.88	56.84	24.36	47.45	43.30	39.81	34.93	37.36	**29.51**	22.54	32.83
+ SC	43.71	48.97	58.68	25.37	50.36	45.42	40.75	35.47	38.54	28.86	26.38	34.00
+ SR	**45.08**	**51.53**	**59.00**	**26.67**	**52.34**	**46.92**	**42.96**	**37.54**	**40.64**	28.34	**29.25**	**35.75**

On the self-guided randomization We then study the choice of layers to perform the self-guided randomization. As shown in Table 7, for both DG settings, we achieve the best performance when randomizing features from both layer1 and layer2, *i.e.*, f^1 and f^2. Besides, for all the test domains, the best performance is always located in the last three rows which also verifies the effectiveness of the self-guided randomization. From the Table 7, we also observe that the croppping operation $\mathcal{C}$ and randomly generated hyper-parameter λ are effective in both generalization scenarios. Our proposed self-guided randomization is also outperform MixStyle [67] on the semantic segmentation task.

Table 7. Ablation study on the choice of layers for the self-guided randomization in our SiamDoGe.

	Trained on Cityscapes						Trained on GTAV					
Variants	G	B	M	S	A	*avg.*	C.	B	M	S	A	*avg.*
w/o randomization	43.71	48.97	58.68	25.37	50.36	45.42	40.75	35.47	38.54	28.86	26.38	34.00
Using MixStyle [67]	44.40	50.87	57.30	24.39	48.85	45.16	40.68	36.00	38.55	27.81	27.69	34.15
w/o $\mathcal{C}$	45.01	51.44	58.62	25.57	52.83	46.69	40.15	37.94	38.18	27.50	28.80	34.51
$\lambda = 0.5$	44.58	50.87	58.67	25.62	51.54	46.26	39.89	38.21	38.24	27.92	28.72	34.60
layer 1 only	**45.53**	**51.85**	**59.13**	25.20	52.01	46.74	41.18	**37.73**	40.34	27.28	**29.60**	35.23
layer 2 only	43.29	50.18	57.49	25.78	49.77	45.30	41.51	37.30	38.65	**29.22**	27.95	34.93
layers 1 & 2	45.08	51.53	59.00	**26.67**	**52.34**	**46.92**	**42.96**	37.54	**40.64**	28.34	29.25	**35.75**

On the sensitivity-guided consistency training Next, we study several variants of the proposed sensitivity-guided consistency training. We can observe from the top part of Table 8 (Rows 1-3) that both the feature-level consistency loss $\mathcal{L}_{f-cs}$ and the prediction-level consistency loss $\mathcal{L}_{p-cs}$ can boost the performance of domain generalization on most test domains and the feature-level one seems even more important. We also find (from Row 4 and Row 5) that the performance of each column is improved except for BDD-100K and ACDC (trained on GTAV), which verifies the importance of sensitivity guidance for consistency training.

Hyper-parameter tuning Finally, we tune the values of the hyper-parameter α and the number of iterations before launching the consistency loss during the training of our full model. The results are reported in the bottom part of Table 8 and we observe that $\alpha = 10$ and $iter = 10$k gives the best performance.

Table 8. Ablation study on each factor of sensitivity-guided consistency loss in our SiamDoGe. "iter" represents the iterations required before launching the sensitivity-guided consistent loss.

Variants				Trained on Cityscapes						Trained on GTAV					
$\mathcal{L}_{f-cs}$	$\mathcal{L}_{p-cs}$	α	iter.	G	B	M	S	A	*avg.*	C	B	M	S	A	*avg.*
		-	-	43.28	49.40	57.93	24.67	47.88	44.63	40.08	34.52	37.94	**28.62**	25.71	33.37
✓		10.0	10k	43.54	48.99	57.40	26.13	48.21	44.85	39.39	34.13	36.44	26.88	27.14	32.80
	✓	10.0	10k	**45.23**	51.21	58.92	26.13	**52.44**	46.79	40.85	37.31	37.99	26.94	27.72	34.12
w/o S	w/o S	10.0	10k	44.26	51.07	57.22	24.55	51.70	45.76	41.86	**39.09**	40.04	26.96	**29.72**	35.53
✓	✓	10.0	10k	45.08	**51.53**	**59.00**	**26.67**	52.34	**46.92**	**42.96**	37.54	**40.64**	28.34	29.25	**35.75**
✓	✓	1.0	10k	44.62	50.17	58.50	25.59	50.41	45.86	41.42	34.96	38.94	27.41	25.61	33.67
✓	✓	50.0	10k	43.84	50.13	56.88	23.75	52.13	45.35	41.07	37.53	40.52	25.52	28.08	34.54
✓	✓	10.0	0	44.06	51.49	58.37	25.18	51.75	46.17	40.28	35.75	38.82	26.43	28.00	33.86
✓	✓	10.0	20k	44.17	51.24	58.01	25.11	51.69	46.04	40.71	36.47	39.86	25.88	27.00	33.98

5 Conclusion

In this paper, we explored a novel domain generalizable semantic segmentation approach with a more controllable domain randomization strategy. The proposed method, "SiamDoGe", is built upon a Siamese network with two branches performing semantic segmentation. It is integrated with two novel designs: one is the self-guided randomization which randomizes each training sample without using auxiliary domain images (different from other existing DR-based alternatives), the other is the sensitivity-guided consistency training which helps learn domain-agnostic features from two views of each training sample. Comprehensive numerical experiments demonstrated that our SiamDoGe generalizes well on several unseen target domains by training on a single domain and achieves a new state-of-the-art performance.

Acknowledgment: Dr. Lili Ju's work is partially supported by U.S. Department of Energy, Office of Advanced Scientific Computing Research through Applied Mathematics program under grant DE-SC0022254. This work used GPUs provided by the NSF MRI-2018966.

References

1. Alhaija, H., Mustikovela, S., Mescheder, L., Geiger, A., Rother, C.: Augmented reality meets computer vision: Efficient data generation for urban driving scenes. IJCV (2018) 1
2. Banerjee, S., Hati, A., Chaudhuri, S., Velmurugan, R.: Cosegnet: Image co-segmentation using a conditional siamese convolutional network. In: IJCAI. pp. 673–679 (2019) 4
3. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.: Fully-convolutional siamese networks for object tracking. In: ECCV. pp. 850–865. Springer (2016) 4
4. Bromley, J., Bentz, J.W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säckinger, E., Shah, R.: Signature verification using a "siamese" time delay neural network. International Journal of Pattern Recognition and Artificial Intelligence **7**(04), 669–688 (1993) 4
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE TPAMI (2018) 1, 8
6. Chen, W., Yu, Z., Wang, Z., Anandkumar, A.: Automated synthetic-to-real generalization. In: Int. Conf. Machine Learning. pp. 1746–1756. PMLR (2020) 4
7. Chen, X., He, K.: Exploring simple siamese representation learning. In: CVPR. pp. 15750–15758 (2021) 3, 4, 8
8. Chen, Y.H., Chen, W.Y., Chen, Y.T., Tsai, B.C., Frank Wang, Y.C., Sun, M.: No more discrimination: Cross city adaptation of road scene segmenters. In: ICCV. pp. 1992–2001 (2017) 2
9. Cheng, S., Zhong, B., Li, G., Liu, X., Tang, Z., Li, X., Wang, J.: Learning to filter: Siamese relation network for robust tracking. In: CVPR. pp. 4421–4431 (2021) 4
10. Choi, S., Jung, S., Yun, H., Kim, J.T., Kim, S., Choo, J.: Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In: CVPR. pp. 11580–11590 (2021) 2, 4, 8, 9, 10, 11, 12
11. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR. pp. 3213–3223 (2016) 1, 9
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255. Ieee (2009) 2, 4
13. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV **88**(2), 303–338 (2010) 1
14. Gan, C., Yang, T., Gong, B.: Learning attributes equals multi-source domain generalization. In: CVPR. pp. 87–97 (2016) 2
15. Gong, R., Li, W., Chen, Y., Gool, L.V.: Dlow: Domain flow for adaptation and generalization. In: CVPR. pp. 2477–2486 (2019) 6
16. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., et al.: Bootstrap your own latent: A new approach to self-supervised learning. arXiv preprint arXiv:2006.07733 (2020) 4
17. Guo, X., Yang, C., Li, B., Yuan, Y.: Metacorrection: Domain-aware meta loss correction for unsupervised domain adaptation in semantic segmentation. In: CVPR. pp. 3927–3936 (2021) 4
18. He, A., Luo, C., Tian, X., Zeng, W.: A twofold siamese network for real-time object tracking. In: CVPR. pp. 4834–4843 (2018) 4

19. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR. pp. 9729–9738 (2020) 4
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) 8
21. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. In: Int. Conf. Machine Learning. pp. 1989–1998. PMLR (2018) 4
22. Hoffman, J., Wang, D., Yu, F., Darrell, T.: Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. arXiv preprint arXiv:1612.02649 (2016) 2, 4
23. Huang, J., Guan, D., Xiao, A., Lu, S.: Fsdr: Frequency space domain randomization for domain generalization. In: CVPR. pp. 6891–6902 (2021) 2, 3, 4
24. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: ICCV. pp. 1501–1510 (2017) 6
25. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Int. Conf. Machine Learning (2015) 4
26. Kim, M., Byun, H.: Learning texture invariant representation for domain adaptation of semantic segmentation. In: CVPR. pp. 12975–12984 (2020) 4
27. Kim, N., Son, T., Lan, C., Zeng, W., Kwak, S.: Wedge: Web-image assisted domain generalization for semantic segmentation. arXiv preprint arXiv:2109.14196 (2021) 2, 3, 4, 9, 10
28. Koch, G., Zemel, R., Salakhutdinov, R., et al.: Siamese neural networks for one-shot image recognition. In: ICML deep learning workshop. vol. 2. Lille (2015) 4
29. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Deeper, broader and artier domain generalization. In: ICCV. pp. 5542–5550 (2017) 2
30. Li, D., Zhang, J., Yang, Y., Liu, C., Song, Y.Z., Hospedales, T.M.: Episodic training for domain generalization. In: ICCV. pp. 1446–1455 (2019) 2
31. Li, H., Pan, S.J., Wang, S., Kot, A.C.: Domain generalization with adversarial feature learning. In: CVPR. pp. 5400–5409 (2018) 2
32. Li, W., Jafari, O.H., Rother, C.: Deep object co-segmentation. In: ACCV. pp. 638–653. Springer (2018) 4
33. Lian, Q., Lv, F., Duan, L., Gong, B.: Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In: ICCV. pp. 6758–6767 (2019) 4
34. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015) 1
35. Luo, Y., Liu, P., Guan, T., Yu, J., Yang, Y.: Significance-aware information bottleneck for domain adaptive semantic segmentation. In: ICCV. pp. 6778–6787 (2019) 4
36. Luo, Y., Liu, P., Guan, T., Yu, J., Yang, Y.: Adversarial style mining for one-shot unsupervised domain adaptation. NeurIPS (2020) 2
37. Ma, N., Zhang, X., Zheng, H.T., Sun, J.: Shufflenet v2: Practical guidelines for efficient cnn architecture design. In: ECCV. pp. 116–131 (2018) 12
38. Mei, K., Zhu, C., Zou, J., Zhang, S.: Instance adaptive self-training for unsupervised domain adaptation. In: ECCV. pp. 415–430. Springer (2020) 4
39. Melas-Kyriazi, L., Manrai, A.K.: Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In: CVPR. pp. 12435–12445 (2021) 4
40. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. IEEE (2016) 1

41. Mousavian, A., Toshev, A., Fišer, M., Košecká, J., Wahid, A., Davidson, J.: Visual representations for semantic target driven navigation. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 8846–8852. IEEE (2019) 1
42. Muandet, K., Balduzzi, D., Schölkopf, B.: Domain generalization via invariant feature representation. In: Int. Conf. Machine Learning. pp. 10–18. PMLR (2013) 2
43. Neuhold, G., Ollmann, T., Rota Bulo, S., Kontschieder, P.: The mapillary vistas dataset for semantic understanding of street scenes. In: ICCV (2017) 9
44. Pan, F., Shin, I., Rameau, F., Lee, S., Kweon, I.S.: Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In: CVPR. pp. 3764–3773 (2020) 2, 4
45. Pan, X., Luo, P., Shi, J., Tang, X.: Two at once: Enhancing learning and generalization capacities via ibn-net. In: ECCV. pp. 464–479 (2018) 2, 4, 9, 10, 11, 12
46. Qiao, F., Zhao, L., Peng, X.: Learning to learn single domain generalization. In: CVPR. pp. 12556–12565 (2020) 2
47. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: ECCV (2016) 9
48. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: CVPR. pp. 3234–3243 (2016) 9
49. Sakaridis, C., Dai, D., Gool, L.V.: Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In: ICCV. pp. 7374–7383 (2019) 4
50. Sakaridis, C., Dai, D., Van Gool, L.: Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. IEEE TPAMI (2020). https://doi.org/10.1109/TPAMI.2020.3045882 4
51. Sakaridis, C., Dai, D., Van Gool, L.: Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. arXiv preprint arXiv:2104.13395 (2021) 4, 9
52. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: CVPR. pp. 4510–4520 (2018) 12
53. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV. pp. 746–760. Springer (2012) 1
54. Tao, R., Gavves, E., Smeulders, A.W.: Siamese instance search for tracking. In: CVPR. pp. 1420–1429 (2016) 4
55. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: CVPR. pp. 7472–7481 (2018) 2, 4
56. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016) 4
57. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: CVPR. pp. 2517–2526 (2019) 2, 4
58. Wang, Z., Luo, Y., Qiu, R., Huang, Z., Baktashmotlagh, M.: Learning to diversify for single domain generalization. In: ICCV. pp. 834–843 (2021) 2
59. Wu, X., Wu, Z., Guo, H., Ju, L., Wang, S.: Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation. In: CVPR. pp. 15769–15778 (2021) 4
60. Yang, Y., Soatto, S.: Fda: Fourier domain adaptation for semantic segmentation. In: CVPR. pp. 4085–4095 (2020) 2, 4

61. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: CVPR (2020) 9
62. Yue, X., Zhang, Y., Zhao, S., Sangiovanni-Vincentelli, A., Keutzer, K., Gong, B.: Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In: CVPR. pp. 2100–2110 (2019) 2, 4, 7, 9, 10
63. Zhang, P., Zhang, B., Zhang, T., Chen, D., Wang, Y., Wen, F.: 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. pp. 12414–12424 (2021) 2, 4
64. Zhang, Z., Peng, H.: Deeper and wider siamese networks for real-time visual tracking. In: CVPR. pp. 4591–4600 (2019) 4
65. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR. pp. 2881–2890 (2017) 1, 8
66. Zhao, L., Liu, T., Peng, X., Metaxas, D.: Maximum-entropy adversarial data augmentation for improved generalization and robustness. In: NeurIPS (2020) 2
67. Zhou, K., Yang, Y., Qiao, Y., Xiang, T.: Domain generalization with mixstyle. In: ICLR (2021) 2, 6, 13
68. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep learning in medical image analysis and multimodal learning for clinical decision support, pp. 3–11. Springer (2018) 1
69. Zou, Y., Yu, Z., Kumar, B., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: Proceedings of the European conference on computer vision (ECCV). pp. 289–305 (2018) 4
70. Zou, Y., Yu, Z., Liu, X., Kumar, B., Wang, J.: Confidence regularized self-training. In: ICCV. pp. 5982–5991 (2019) 4