Style Projected Clustering for Domain Generalized Semantic Segmentation

Wei Huang[1*] Chang Chen[2†] Yong Li[2] Jiacheng Li[1]
Cheng Li[2] Fenglong Song[2] Youliang Yan[2] Zhiwei Xiong[1]
[1]University of Science and Technology of China [2]Huawei Noah's Ark Lab
{weih527,jclee}@mail.ustc.edu.cn, zwxiong@ustc.edu.cn,
{chenchang25,liyong156,licheng89,songfenglong,yanyouliang}@huawei.com

Abstract

Existing semantic segmentation methods improve generalization capability, by regularizing various images to a canonical feature space. While this process contributes to generalization, it weakens the representation inevitably. In contrast to existing methods, we instead utilize the difference between images to build a better representation space, where the distinct style features are extracted and stored as the bases of representation. Then, the generalization to unseen image styles is achieved by projecting features to this known space. Specifically, we realize the style projection as a weighted combination of stored bases, where the similarity distances are adopted as the weighting factors. Based on the same concept, we extend this process to the decision part of model and promote the generalization of semantic prediction. By measuring the similarity distances to semantic bases (i.e., prototypes), we replace the common deterministic prediction with semantic clustering. Comprehensive experiments demonstrate the advantage of proposed method to the state of the art, up to 3.6% mIoU improvement in average on unseen scenarios. Code and models are available at https://gitee.com/mindspore/models/tree/master/research/cv/SPC-Net.

1. Introduction

Domain generalization methods aim to promote the performance of model (trained on source datasets), when applying it to *unseen* scenarios (target domains) [9, 19, 29, 36, 62, 74, 75]. Recently, domain generalization for semantic segmentation (DGSS) has attracted increasingly more attention due to the rise of safety-critical applications, such as autonomous driving [3, 12, 22, 45].

Existing DGSS methods improve the pixel-wise generalization performance by learning domain-agnostic representations [5, 16, 20, 25, 40, 42, 66, 72]. Researches in this line share the similar goal in general, that is to capture the domain-invariant characteristics of object contents, and eliminates the domain-specific ones (*i.e.*, image styles). As two representatives, Instance Normalization (IN) [56] and Instance Whitening (IW) [17] regularize image features from different domains to a canonical space, as illustrated in Fig. 1(a) and 1(b). Specifically, IN achieves center-level feature alignment via channel-wise feature normalization [33,40], and IW realizes uniform feature distribution by removing linear correlation between channels [5,41]. Moreover, the combination of these two methods is proposed in [42] for a better generalization, as shown in Fig. 1(c).

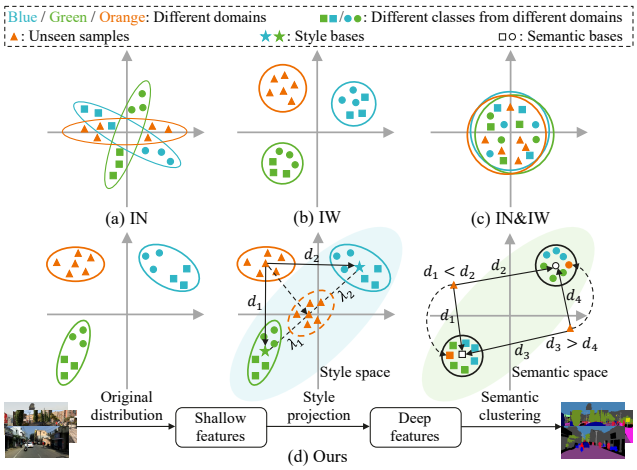


Figure 1. Illustration of instance normalization/whitening (IN/IW) [5, 20, 40] and our proposed style projected clustering method. IN and IW regularize image features from different domains to a canonical space (a-c). Our method builds style and semantic representation spaces based on the data from known domains (d).

Nevertheless, feature regularization inevitably weakens the representation capability, as a part of feature information is eliminated. Theoretically, it works under a strong assumption that the eliminated information is strictly the domain-specific ones. Yet in practice, the perfect disentanglement between image style and content is difficult to

*This work was done during W. Huang's internship at Noah's Ark Lab.
†Corresponding author

achieve. It means that a part of content features will also be eliminated in the process of feature regularization, and thus degrades the segmentation performance.

Instead of seeking common ground by feature regularization, we aim to address DGSS in a different way. In this paper, we propose *style projection* as an alternative, which utilizes the features from different domains as bases to build a better representation space, as shown in Fig. 1(d). The motivation of style projection comes from a basic concept of generalization, that is to represent unseen data based on the known ones. Specifically, following the common practice, we adopt the statistics (*i.e.*, mean and variance) of features in channel dimension to represent image styles. The image styles from source domains are iteratively extracted and stored as the bases of representation. Then, we project the style of given unseen images into this representation space to promote generalization. This projection process is implemented as a weighted combination of stored style bases, where the similarity distance between styles are adopted as the weighting factors, *i.e.*, λ_1 and λ_2 shown in Fig. 1(d).

Based on the projected style features, we further devise the decision part of model, which is elaborated for semantic segmentation. Typically, existing methods learn a parametric function to map pixel-wise features to semantic predictions. We replace this deterministic prediction with *semantic clustering*, where the class of each pixel is predicted by the minimal similarity distance to semantic bases, as shown in Fig. 1(d). Notably, it follows the same concept of style projection, that is to predict unseen data based on the known ones. More concretely, to facilitate the performance of semantic clustering, we propose a variant of contrastive loss to align the semantic bases of same classes and enhance discriminability between different classes.

We conduct comprehensive experiments on single- and multi-source settings to demonstrate the superior generalization of our method over existing DGSS methods. In addition, we visually analyze the effective representation of our proposed method for unseen images in both style and semantic spaces.

Contributions of this paper are summarized as follows:

- Beyond existing feature regularization methods, we propose style projected clustering, pointing out a new avenue to address DGSS.
- We propose style projection, which projects unseen styles into the style representation space built on known domains for a better representation.
- We propose semantic clustering to predict the class of each pixel in unseen images by the similarity distance to semantic bases, which further improves the generalization capability for unseen domains.
- Our proposed method outperforms the current state of the arts on multiple DGSS benchmarks.

2. Related Work

Domain adaptation and generalization. To reduce the burden of pixel-wise annotations on target domains, domain adaptation (DA) technologies are proposed to narrow the domain gap between source and target domains via image translation [14, 24, 37], feature alignment [55, 60, 61], self-training [2, 39, 77] and meta-learning [13, 34] strategies. However, these DA methods require the access of data on target domains. Domain generalization (DG) aims to address a more practical problem where the target domain cannot be accessed. Numerous DG works have been proposed for image classification via style augmentation [19, 59, 68, 75], domain alignment [29, 31], feature disentanglement [27, 44] and meta-learning [9, 26, 28].

Domain generalization for semantic segmentation. Similar to image classification, DG for semantic segmentation (DGSS) methods are proposed to learn domain-agnostic representations, including style augmentation [16, 25, 43, 72], feature normalization/whitening [5, 40, 42, 66] and meta-learning [20]. To avoid overfitting on source domains, DRPC [72] and FSDR [16] adopt style augmentations in the image space to extend the number of source samples, while WildNet [25] realizes it in the feature space with the aid of ImageNet [8]. Alternatively, normalization and whitening are investigated to achieve distribution alignment between different domains. IBN-Net [40] and RobustNet [5] adopt instance normalization and whitening, respectively, to eliminate the specific style information of each domain. Furthermore, SAN-SAW [42] proposes semantic-aware instance normalization and whitening to enhance the distinguishability between classes. In addition, PintheMem [20] combines the memory-guided network with the meta-learning strategy and obtains competitive performances. Different from these DGSS methods, our method embraces the differences from multiple known domains and takes advantage of their diversity to build a better representation space, realizing the representation of unseen images by the known data.

Prototype learning. Inspired by the cognitive psychology that human use the knowledge learned in the past to judge the class of unknown things [51,69], prototype-based classification methods have attracted increasing attention, where the class of unknown images is determined by its nearest neighbors in the feature space [7, 10]. Owing to its excellent interpretability and generalizability, prototype learning shows good potential in many fields, such as few-shot learning [1, 52], zero-shot learning [67, 71], unsupervised learning [30, 65]. Recently, prototype learning is also introduced in the dense prediction task, including supervised [76], few-shot [54,63] and domain adaptive [53,73] semantic segmentation. To facilitate the learning of prototypes, metric learning [23,50,64] is often adopted to pull samples belonging to the same class together and push those of different classes away from each other in the embedding (*i.e.*, feature) space.

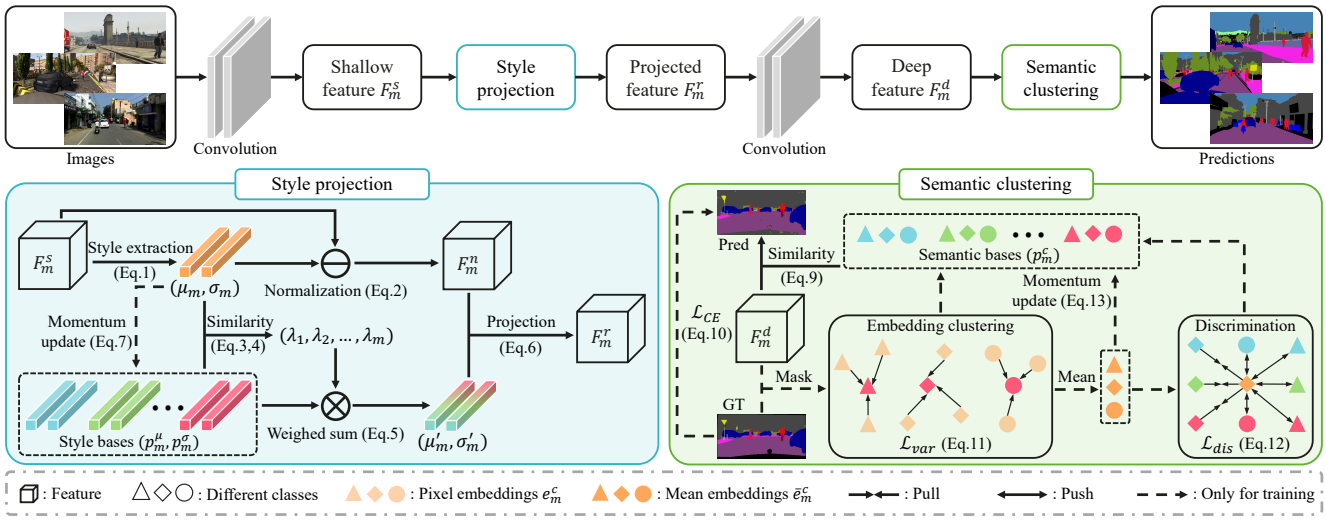


Figure 2. The framework of style projected clustering, which consists of two components, *i.e.*, style projection and semantic clustering. We iteratively extract the style and semantic information of seen domains as style bases (p_m^μ, p_m^σ) and semantic bases p_m^c. In style projection, we first calculate the similarity between the unseen style (μ_m, σ_m) from the shallow feature F_m^s and style bases (p_m^μ, p_m^σ) as weighted factors λ_m. Then, the weighted combination of style bases (μ'_m, σ'_m) is projected on F_m^n to obtain the projected feature F_m^r. In semantic clustering, we calculate the similarity between pixel embeddings in the deep feature F_m^d and semantic bases p_m^c. Then, the class of each pixel is determined by the nearest semantic base. During the training phase, the cross-entropy loss $\mathcal{L}_{CE}$, variance loss $\mathcal{L}_{var}$ and discrimination loss $\mathcal{L}_{dis}$ are adopted to supervise the learning of style and semantic bases.

Similar to these methods, we adopt the form of prototypes (*i.e.* bases) to represent semantics. Yet these semantic bases are learned in a different way to facilitate domain generalization, by using a new variant of contrastive loss.

3. Style Projected Clustering

The overall architecture of our proposed method is depicted in Fig. 2, which consists of two components, *i.e.*, style projection and semantic clustering. In style projection, we project the unseen style into the style representation space built on style bases, according to the similarity between the unseen style and style bases. In semantic clustering, we estimate the similarity between pixel embeddings and semantic bases (*i.e.*, prototypes) to determine the class of pixels in unseen images by the nearest semantic base.

3.1. Problem Formulation

In the domain generalized semantic segmentation problem, we are given M source domains $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_M\}$ that are from multiple datasets with different data distributions. The m-th source domain $\mathcal{S}_m$ can be represented as $\mathcal{S}_m = \{(x_m, y_m)\}$, where $x_m \in \mathbb{R}^{H\times W\times 3}$ is an image from the m-th source domain, $y_m \in \mathbb{R}^{H\times W\times C}$ is the corresponding pixel-wise label, C is the number of semantic classes, H and W are the height and width of the image x_m, respectively. In this work, our goal is to train a semantic segmentation model ϕ to obtain the best generalization performance on multiple target domains $\mathcal{T}$ which cannot be accessed during the training phase.

3.2. Style Projection

The style difference of images is the main factor leading to the domain shift, which limits the generalization ability of the learned model. Pioneering works [11,18,40,75] have demonstrated that the feature distribution shift caused by style differences lies mainly in shallow layers of networks. It also shows that the shallow feature distribution of networks can reflect the style information of the input image x_m. Thus, existing works always adopt the channel-wise mean and variance of the shallow feature to represent the style distribution of x_m [18,25]. Following these works, let $F_m^s \in \mathbb{R}^{D\times H_s\times W_s}$ be the shallow feature of x_m from the network ϕ, where D denotes the number of channels. The channel-wise mean $\mu_m \in \mathbb{R}^D$ and variance $\sigma_m \in \mathbb{R}^D$ of the feature F_m^s can be calculated as follows:

$$\begin{aligned}\mu_m &= \frac{1}{H_s W_s}\sum_{h=1}^{H_s}\sum_{w=1}^{W_s} F_m^s, \\ \sigma_m &= \sqrt{\frac{1}{H_s W_s}\sum_{h=1}^{H_s}\sum_{w=1}^{W_s}(F_m^s - \mu_m)^2}.\end{aligned} \tag{1}$$

To eliminate the specific style information of images, instance normalization [40] is adopted to standardize the feature F_m^s to a standard distribution (*i.e.*, zeros mean and one standard deviation) as follows:

$$F_m^n = \frac{F_m^s - \mu_m}{\sigma_m + \epsilon}, \tag{2}$$

where F_m^n stands for the normalized feature, and ϵ is a small value to avoid division by zero.

Although instance normalization achieves to remove the specific style information of images, it also eliminates the natural differences between domains, which weakens the representation for target domains and produces limited generalization performance. Therefore, to preserve the specific style information of each domain, we propose style bases $P_{sty} = \{(p_m^{\mu}, p_m^{\sigma})\}_{m=1}^{M}$ to store the style information of source domains, and then leverage the preserved style bases P_{sty} to build a style representation space, realizing the projection of unseen style, as shown in Fig. 2. Specifically, we first leverage Wasserstein distance [57] to estimate the style distribution discrepancy between the input image x_m and the m-th style bases $(p_m^{\mu}, p_m^{\sigma})$ as follows:

$$d_m = ||\mu_m - p_m^{\mu}||_2^2 + ({\sigma_m}^2 + {p_m^{\sigma}}^2 - 2\sigma_m p_m^{\sigma}), \quad (3)$$

where d_m denotes the distribution distance between the current image x_m and the m-th source domain. Then, we use the reciprocal of d_m to characterize the similarity between x_m and m-th style bases as follows:

$$\lambda_m = \frac{exp(1/(1+d_m))}{\sum_{m=1}^{M} exp(1/(1+d_m))}, \quad (4)$$

where the softmax operation is utilized to make the sum of $\lambda = \{\lambda_m | m = 1, 2, ..., M\}$ equal to 1. Based on the estimated similarity λ, we can obtain the projected style $({\mu_m}', {\sigma_m}')$ by the weighted sum of style bases as follows:

$${\mu_m}' = \sum_{m=1}^{M} \lambda_m p_m^{\mu}, \ {\sigma_m}' = \sum_{m=1}^{M} \lambda_m p_m^{\sigma}. \quad (5)$$

Finally, following previous works [11, 18, 19, 25], we inject the projected style $({\mu_m}', {\sigma_m}')$ into the normalized feature F_m^n to obtain the projected feature as follows:

$$F_m^r = {\sigma_m}' F_m^n + {\mu_m}'. \quad (6)$$

During the training phase, we adopt the momentum update strategy to achieve the online collection of style information as follows:

$$\begin{aligned} p_m^{\mu} &= \alpha p_m^{\mu} + (1-\alpha)\mu_m, \\ p_m^{\sigma} &= \alpha p_m^{\sigma} + (1-\alpha)\sigma_m, \end{aligned} \quad (7)$$

where $\alpha \in [0, 1]$ is a momentum coefficient. In addition, we randomly initialize P_{sty} to start training, where p_m^{μ} and p_m^{σ} are initialized with zero-mean and one-mean distribution, respectively. By Eq. 7, we realize the style statistic of source domains and store it as style bases efficiently.

After style projection, the projected feature F_m^r is input into the next layer of the network ϕ. Our style projection is designed as a plug-and-play module that can be applied behind any network layer. However, as the layer is deeper, the style information loosens while the semantic information plays a more important role. Thus, in this work, style projection is only used in the first two layers of ϕ to obtain the best generalization performance.

3.3. Semantic Clustering

To obtain the final pixel-wise predictions, we further propose semantic clustering on the deep feature extracted by the network ϕ. Let $F_m^d \in \mathbb{R}^{D \times H_d \times W_d}$ be the deep feature of the input image x_m from ϕ. Existing DGSS methods generically apply a learnable segmentation classifier ϕ_{cls} on F_m^d for the dense prediction. However, the parameters of ϕ_{cls} is learned on the deep features of source domains $\mathcal{S}$, and thus its generalization ability on target domain $\mathcal{T}$ is limited. In addition, the semantic information between different domains is implicitly encoded in the same parameter space, which causes the specific semantic information of domains to be eliminated.

Based on the concept of style bases, we introduce semantic bases $P_{sem} = \{p_m^c\}_{c,m=1}^{C,M}$ to preserve the semantic information of each domain and each class, where $p_m^c \in \mathbb{R}^D$ is the cluster center of training pixel embeddings belonging to the c-th class from the m-th source domain in the feature space. Following the prototype theory [7, 10, 76], the class of each pixel embedding $e \in F_m^d$ can be determined by its nearest semantic bases as follows:

$$c(e) = c^*, \text{with}(c^*, m^*) = \text{argmin}_{c,m}\{d_m^c\}_{c,m=1}^{C,M}, \quad (8)$$

where $d_m^c = -cos(e, p_m^c)$ is the negative cosine distance used to estimate the similarity between the current embedding e and semantic bases p_m^c. In this work, the pixel embedding e and semantic bases p_m^c are both l_2-normalized. Therefore, the similarity distance can be simply formulated as $d_m^c = -ep_m^c$. Different from the learnable segmentation classifier ϕ_{cls}, P_{sem} not only explicitly captures characteristic properties of each class from each domain, but also determines the class of pixels in unseen images without introducing extra learnable parameters.

To facilitate the training of the network ϕ during the training phase, we estimate the probability value of pixel embedding e belonging to class c as follows:

$$v(c|e) = \frac{exp(-d^c)}{\sum_{c=1}^{C} exp(-d^c)}, \quad (9)$$

where $d^c = \min_m\{d_m^c\}_{m=1}^{M}$ denotes the similarity between e and its closet semantic base belonging to class c. Then, we adopt the standard cross-entropy loss to supervise the training of the network ϕ as follows:

$$\mathcal{L}_{CE} = -\frac{1}{H_d W_d} \sum_{h=1}^{H_d} \sum_{w=1}^{W_d} \sum_{c=1}^{C} y_m log(v(c|e)), \quad (10)$$

where y_m is the pixel-wise label corresponding to the input image x_m.

However, the naive cross-entropy loss only optimizes the relative relations between intra-class and inter-class distance, which ignores the absolute distance constraint between pixel embeddings and semantic bases. That is to say,

we expect that the pixel embedding belonging to class c is closer to the c-th semantic base and is farther away from the semantic bases belonging to other classes. Inspired by metric learning [21, 23], we further propose variance and discrimination terms as two extra training objectives. The former is an intra-class cluster that pulls the pixel embedding e_m^c belonging to class c from the m-th source domain towards the semantic bases p_m^c:

$$\mathcal{L}_{var} = \frac{1}{MC}\sum_{m=1}^{M}\sum_{c=1}^{C}(1 - e_m^c p_m^c)^2. \tag{11}$$

The latter is designed in a contrastive learning way which encourages the current cluster center $\bar{e}_m^c$ is closer to the c-th semantic bases p_{c+} (*i.e.*, positive keys) and to be far away from semantic bases belonging to other class p_{c-} (*i.e.*, negative keys):

$$\mathcal{L}_{dis} = \frac{1}{M}\sum_{p_{c+}} -log\frac{exp(\bar{e}_m^c p_{c+}/\tau)}{exp(\bar{e}_m^c p_{c+}/\tau) + \sum_{p_{c-}} exp(\bar{e}_m^c p_{c-}/\tau)}, \tag{12}$$

where $\bar{e}_m^c$ is the cluster center (*i.e.*, mean embedding) of pixel embedding e_m^c in the current feature F_m^d, and τ is a temperature hyper-parameter. By Eq. 12, we realize the alignment of semantic bases belonging to the same class c from different domains. Different from existing pixel-wise contrastive learning paradigm [64], the positive and negative samples in Eq. 12 are semantic bases rather than pixel embeddings. Thus, we don't need to construct a memory bank to store sufficient embedding samples, which also significantly reduces the computational cost.

To achieve the online collocation of semantic information from source domains, we adopt the same momentum update strategy to update semantic bases P_{sem} as follows:

$$p_m^c = \alpha p_m^c + (1 - \alpha)\bar{e}_m^c, \tag{13}$$

where α the momentum coefficient. Like style bases, we also randomly initialize the semantic bases P_{sem} with zero-mean distribution to start our training.

3.4. Training and Inference

During the training phase, we combine above three loss terms for the end-to-end training as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \beta\mathcal{L}_{var} + \gamma\mathcal{L}_{dis}, \tag{14}$$

where β and γ are weighting coefficients to balance these three terms. For each training iteration, in addition to the parameter update of the network ϕ, the style and semantic bases are also updated online by Eq. 7 and Eq. 13.

During the inference phase, we leverage Eq. 8 to obtain final pixel-wise predictions by the nonparametric cluster of pixel embeddings outputted from the learned network ϕ.

4. Experiments

4.1. Datasets

Synthetic datasets. GTAV [47] contains 24966 images with a resolution of 1914×1052 captured from the GTA-V game engine. Synthia [48] contains 9400 images with a resolution of 1280×760 generated from virtual urban scenes.
Real-world datasets. IDD [58] contains 10004 images with an average resolution of 1678×968 captured from Indian roads. Cityscapes [6] contains 5000 fine annotated images with a resolution of 2048×1024 captured from 50 different cities primarily in Germany. BDD100K [70] contains 10000 image with a resolution of 1280×720 captures from different locations in US. Mapillary [38] contains 25000 images with an average resolution of 1920×1080 captured from all around the world.

4.2. Implementation Details

Following the previous work [5], we adopt DeepLabV3+ [4] with ResNet-50, ResNet-101 [15], MobileNetV2 [49] and ShuffleNetV2 [35] backbones as our segmentation networks, where all backbones are pre-trained on ImageNet [8]. During the training phase, we adopt the SGD optimizer [46] with a momentum of 0.9 and weight decay of $5e-4$. The initial learning rate is set to 0.01 and is decreased using the polynomial scheduling with a power of 0.9. We train all models for 40K iterations, except for the three-source setting, the model is trained for 100K iterations. In addition to some common data augmentations used in [5], we adopt extra strong style augmentations to enrich the style information of urban-scene images [32], which aims to enhance the proposed style projection ability in networks. More details can be found in our supplementary materials.

4.3. Results

Comparison methods. We extensively compare our proposed method against existing DGSS methods, which can be classified into three groups, including style augmentation (WildNet [25]), feature normalization/whitening (IBN-Net [40], RobustNet [5] and SAN-SAW [42]), and meta-learning (MLDG [26] and PintheMem [20]). Since SAN-SAW [42] and WildNet [25] are only implemented on the single-source setting in their paper, we reproduce them on the multi-source setting to make a comparison. In particular, WildNet [25] utilizes the external dataset (*i.e.*, ImageNet) to extend the style and content information of source domains. Thus, we re-implement it by replacing the external dataset with the source dataset for a fair comparison, which is marked with * in our tables.
Multi-source setting. To demonstrate the effectiveness of our proposed method, we first conduct contrast experiments on the multi-source DGSS setting, where multiple source domains can be efficiently used to build a diverse

Methods	Publication	Cityscapes	BDD100K	Mapillary	Avg.-$\mathcal{T}$	GTAV	Synthia	Avg.-$\mathcal{S}$	Avg.-$\mathcal{A}$
Baseline†	-	35.46	25.09	31.94	30.83	68.48	67.99	68.24	45.79
IBN-Net† [40]	ECCV 2018	35.55	32.18	38.09	35.27	69.72	66.90	68.31	48.49
RobustNet† [5]	CVPR 2021	37.69	34.09	38.49	36.76	68.26	68.77	68.52	49.46
Baseline‡	-	33.42	29.07	32.19	31.56	69.63	63.93	66.78	45.65
MLDG‡ [26]	AAAI 2018	38.84	31.95	35.60	35.46	64.61	51.69	58.15	44.54
PintheMem‡ [20]	CVPR 2022	44.51	38.07	42.70	41.76	65.85	54.49	60.17	49.12
Baseline	-	36.03	28.15	32.61	32.26	69.30	67.61	68.46	46.65
SAN-SAW [42]	CVPR 2022	42.13	37.74	42.91	40.93	63.98	62.58	63.28	49.87
WildNet [25]	CVPR 2022	43.65	39.90	43.28	42.28	68.05	63.98	66.02	51.77
WildNet* [25]	CVPR 2022	39.33	34.76	41.06	38.38	69.70	62.11	65.91	49.39
Ours	-	**46.36**	**43.18**	**48.23**	**45.92**	**72.46**	**74.87**	**73.67**	**57.02**

Table 1. **Source (G+S) → Target (C, B, M):** Mean IoU(%) comparison of existing DGSS methods, where all networks with the ResNet-50 backbone are trained with two synthetic (GTAV, Synthia) datasets. The best and second best results are **highlighted** and underlined. Avg.-$\mathcal{T}$, Avg.-$\mathcal{S}$ and Avg.-$\mathcal{A}$ denote the average results on target, source and all domains, respectively. Results with the † and ‡ sign are from [5] and [20], respectively. * indicates that we replace the external dataset (*i.e.*, ImageNet) used in WildNet [25] with the source dataset for a fair comparison.

Methods	Cityscapes	BDD100K	Mapillary	Avg.-$\mathcal{T}$	GTAV	Synthia	IDD	Avg.-$\mathcal{S}$	Avg.-$\mathcal{A}$
Baseline‡	52.51	47.47	54.70	51.56	70.31	67.13	71.56	69.67	60.61
IBN-Net‡ [40]	54.39	48.91	56.06	53.12	70.73	63.68	71.02	68.48	60.80
RobustNet‡ [5]	54.70	49.00	56.90	53.53	70.06	66.40	71.02	69.16	61.35
MLDG‡ [26]	54.76	48.52	55.94	53.07	69.53	59.79	67.73	65.68	59.38
PintheMem‡ [20]	56.57	50.18	58.31	55.02	69.99	62.99	67.58	66.85	60.94
Baseline	54.16	46.24	55.57	51.99	68.35	65.12	70.07	67.85	59.92
SAN-SAW [42]	54.89	46.50	56.38	52.59	64.49	64.76	66.37	65.21	58.90
WildNet [25]	55.58	50.31	57.93	54.61	67.65	61.35	70.07	66.36	60.48
WildNet* [25]	53.61	48.92	56.18	52.90	70.98	59.69	64.52	65.06	58.98
Ours	**57.91**	**53.26**	**61.61**	**57.59**	**74.64**	**78.35**	**76.07**	**76.35**	**66.97**

Table 2. **Source (G+S+I) → Target (C, B, M):** Mean IoU(%) comparison of existing DGSS methods, where all networks with the ResNet-50 backbone are trained with two synthetic (GTAV, Synthia) and one real (IDD) datasets. Results with the ‡ sign are from [20].

representation space. As listed in Table 1, we quantitatively compare our results with existing DGSS methods on both target and source datasets, where all networks with the ResNet-50 backbone are trained with two synthetic datasets (*i.e.*, GTAV and Synthia). Remarkably, compared with the state-of-the-art method (*i.e.*, WildNet [20]), our method not only shows superior generalization capability on target datasets (up to 3.6% mIoU in average), but also significantly improve the performance on source datasets (up to 7.6% mIoU), which demonstrates our method can enhance the representation ability of the learned model on both source and target domains. Furthermore, We provide visual prediction results for qualitative comparisons as shown in Fig. 3. Our method obtains the best visual results on different target datasets. Following [20], we add one real dataset (*i.e.*, IDD) to source domains to further verify the superiority of our method on more source datasets. As listed in Table 2, our method also outperforms existing methods on both source and target domains by a large margin.

Single-source setting. We further implement our method in the single-source setting to make a comprehensive comparison, where all network with the ResNet-50 backbone are trained with one synthetic (*i.e.*, GTAV) dataset. As listed in Table 3, our method shows superior generalization performances over existing DGSS methods. Compared with the naive baseline, our method brings approximately 14% mIoU gains in average on target datasets.

Different backbones. To demonstrate the wide applicability of our method, we compare our results with classic DGSS methods (*i.e.*, IBN-Net [40] and RobustNet [5]) with different backbones. As listed in Table 4, our method shows superior performances on both large (*i.e.*, ResNet-101) and lightweight (*i.e.*, MobileNet and ShuffleNet) backbones.

4.4. Ablation Studies

We conduct comprehensive ablation studies with the ResNet-50 backbone on two source domains (*i.e.*, GTAV and Synthia) as following.

Proposed strategies. As listed in Table 5, our method shows the best generalization capability when two strategies

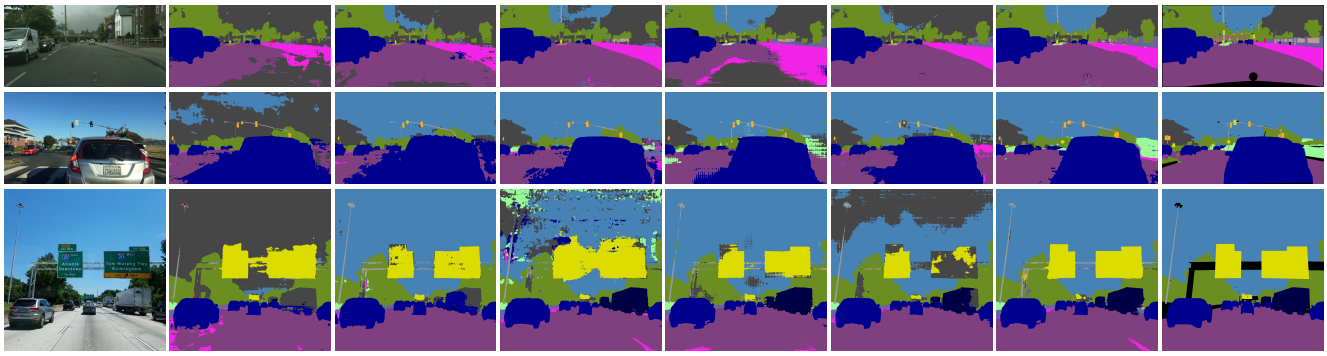


Figure 3. **Source (G+S) → Target (C, B, M):** Visualization comparison with existing DGSS methods on three different target domains.

Methods	C	B	M	Avg.-$\mathcal{T}$
Baseline	28.95	25.14	28.18	27.42
IBN-Net [40]	33.85	32.30	37.75	34.63
RobustNet [5]	36.58	35.20	40.33	37.37
Baseline	31.60	26.70	29.00	29.10
MLDG [26]	36.70	32.10	32.20	33.67
PintheMem [20]	41.00	34.60	37.40	37.67
Baseline	29.32	25.71	28.33	27.79
SAN-SAW [42]	39.75	37.34	41.86	39.65
Baseline	35.16	29.71	31.29	32.05
WildNet [25]	**44.62**	38.42	**46.09**	43.04
Baseline	32.01	26.04	29.35	29.13
WildNet* [25]	40.10	34.82	39.38	38.10
Ours	44.10	**40.46**	45.51	**43.36**

Table 3. **Source (G) → Target (C, B, M):** Mean IoU(%) comparison of existing DGSS methods, where all networks with the ResNet-50 backbone are trained with the one synthetic (GTAV) dataset. * indicates that we replace the external dataset (*i.e.*, ImageNet) used in WildNet [25] with the source dataset for a fair comparison.

	Methods	C	B	M	Avg.-$\mathcal{T}$
MobileNet	Baseline	29.16	20.27	27.19	25.24
	IBN-Net [40]	29.58	26.02	26.32	27.31
	RobustNet [5]	30.67	25.02	28.27	27.99
	Ours	**39.88**	**34.83**	**38.91**	**37.87**
ShuffleNet	Baseline	29.48	26.27	31.35	29.03
	IBN-Net [40]	32.61	29.55	33.20	31.79
	RobustNet [5]	33.15	31.98	34.85	33.33
	Ours	**38.97**	**34.62**	**39.66**	**37.75**
ResNet-101	Baseline	34.71	29.32	37.74	33.92
	IBN-Net [40]	39.18	34.00	39.32	37.50
	RobustNet [5]	39.96	34.94	41.72	38.87
	Ours	**47.93**	**43.62**	**48.79**	**46.78**

Table 4. **Source (G+S) → Target (C, B, M):** Mean IoU(%) comparison of existing DGSS methods with different backbones.

Sty.-Pro.	Sem.-Clu.	C	B	M	Avg.-$\mathcal{T}$
		36.03	28.15	32.61	32.26
✔		44.87	42.42	46.37	44.55
	✔	39.01	30.60	35.19	34.93
✔	✔	**46.36**	**43.18**	**48.23**	**45.92**

Table 5. Ablation results for each strategy used in our method. Sty.-Pro. and Sem.-Clu. indicate style projection and semantic clustering, respectively.

Methods	C	B	M	Avg.-$\mathcal{T}$
Normalization	43.83	40.95	44.92	43.23
Substitution	45.00	42.79	45.16	44.32
Ours	**46.36**	**43.18**	**48.23**	**45.92**

Table 6. Ablation results for different ways of style projection.

are adopted at the same time. Remarkably, compared with the first and second lines, we can find that style projection can approximately bring 12% mIoU gains in average over the baseline, which fully demonstrates its effectiveness for the generalization on unseen domains.

Different ways of style projection. As listed in Table 6, we investigate the effect of different ways of style projection. There are two intuitive ways as follows. One way is using the naive instance normalization to project images from different domains into a normalized feature space (*i.e.*, Normalization). The other way is using the extracted style bases to directly substitute the unseen style (*i.e.*, Substitution). We can find that the weighted combination of style bases can effectively enhance the representation of unseen style, producing better generalization on unseen domains.

Loss terms. As listed in Table 7, we conduct ablation experiments to demonstrate the effectiveness of two complementary loss functions in Eq. 11 and Eq. 12. Compared with the naive cross-entropy loss, adding any complementary loss can bring the performance gain, which verifies each of them can effectively supplement the main loss $\mathcal{L}_{CE}$.

$\mathcal{L}_{CE}$	$\mathcal{L}_{var}$	$\mathcal{L}_{dis}$	C	B	M	Avg.-$\mathcal{T}$
✔			44.00	41.82	45.97	43.93
✔	✔		45.57	42.78	46.61	44.99
✔		✔	45.92	42.42	47.08	45.14
✔	✔	✔	**46.36**	**43.18**	**48.23**	**45.92**

Table 7. Ablation results for each loss term.

Methods	# of Params	GFLOPs	Time (ms)
Baseline	45.08M	277.77	7.82
IBN-Net [40]	45.08M	277.82	8.74
RobustNet [5]	45.08M	277.78	9.48
MLDG [26]	45.08M	277.77	9.67
PintheMem [20]	45.28M	278.31	11.64
SAN-SAW [42]	25.63M	421.86	57.58
WildNet [25]	45.21M	277.16	8.61
Ours	45.22M	286.09	9.98

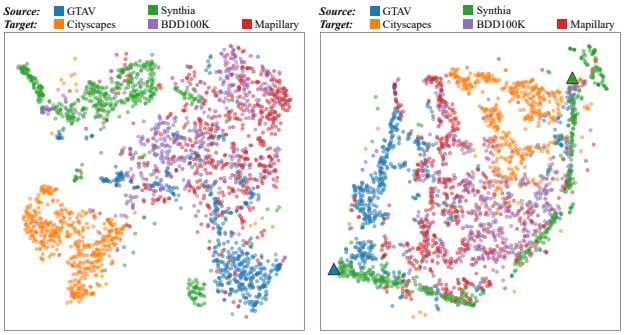
Table 8. Comparison of computational cost. Tested with the image size of 2048×1024 on one NVIDIA Tesla V100 GPU. We average the inference time over 500 trials.

5. Discussion and Analysis

Distribution analysis. We adopt the t-SNE visualization tool to analyze the effectiveness of our proposed style projection and semantic clustering strategies. As shown in Fig 4, we show the variations of style distribution between different domains before and after style projection. We can find that the style distribution of different domains is well separated before style projection (Fig. 4(a)), while their style distribution is approximately constrained between two style bases after style projection (Fig. 4(b)), which demonstrates style projection successfully projects unseen styles into the style representation space built on style bases.

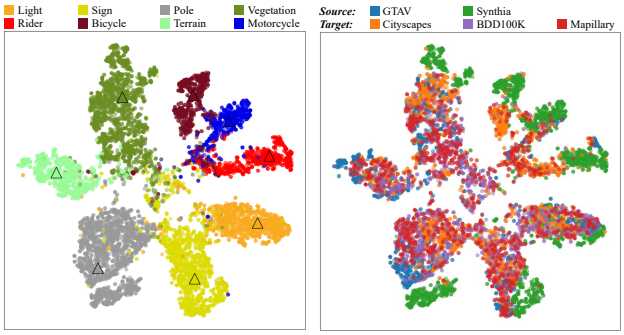
Furthermore, we visualize the semantic distribution between different classes and domains as shown in Fig 5. From Fig. 5(a), we can find that pixel samples belonging to the same class are well clustered while those belonging to different classes are well separated. In addition, the preserved semantic bases are approximately located in the cluster center of pixel samples. From Fig. 5(b), we can find that these pixel samples from different domains are well clustered according to their classes, which demonstrates our semantic clustering successfully achieves the class prediction between different domains by the preserved semantic bases.

Complexity of networks. As listed in Table 8, we compare the number of parameters and computational cost with existing DGSS methods. Since we need to store style and semantic bases and estimate the similarity between them and unseen images, the number of parameters and computational cost in our method are slightly higher than the naive baseline. However, our inference time is competitive to exiting DGSS methods due to the efficient implementation of distance measures by matrix multiplications.



(a) Before projection (b) After projection

Figure 4. t-SNE visualization of style statistics between different domains before (a) and after (b) style projection, where the style statistics (concatenation of mean and variance) is computed from the first layer's feature map of the ResNet-50 trained on two synthetic datasets. Triangles indicate the preserved style bases.



(a) Class distribution (b) Domain distribution

Figure 5. t-SNE visualization of semantic statistics between different classes (a) and domains (b), where the semantic statistics is computed from the last layer's feature map. Triangles indicate the preserved semantic bases.

6. Conclusion

In this paper, we propose a novel style projected clustering method for domain generalized semantic segmentation, which achieves the style and semantic representation of unseen images based on known data. In particular, style projection projects arbitrary unseen styles into the style representation space of source domains and achieves the retention of specific style information between different domains. Semantic clustering predicts the class of each pixel by the minimal similarity distance to semantic bases, which realizes the semantic representation for unseen images and promotes the generalization ability. Through the evaluation on multiple urban-scene datasets, we demonstrate the superior generalization performance of our proposed method over existing DGSS methods.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grants 62131003 and 62021001. And we gratefully acknowledge the support of MindSpore (https://www.mindspore.cn/).

References

[1] Kelsey Allen, Evan Shelhamer, Hanul Shin, and Joshua Tenenbaum. Infinite mixture prototypes for few-shot learning. In *ICML*, 2019. 2

[2] Nikita Araslanov and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *CVPR*, 2021. 2

[3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 1

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 5

[5] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *CVPR*, 2021. 1, 2, 5, 6, 7, 8

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 5

[7] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967. 2, 4

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2, 5

[9] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *NeurIPS*, 2019. 1, 2

[10] Salvador Garcia, Joaquin Derrac, Jose Cano, and Francisco Herrera. Prototype selection for nearest neighbor classification: Taxonomy and empirical study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):417–435, 2012. 2, 4

[11] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016. 3, 4

[12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 1

[13] Xiaoqing Guo, Chen Yang, Baopu Li, and Yixuan Yuan. Metacorrection: Domain-aware meta loss correction for unsupervised domain adaptation in semantic segmentation. In *CVPR*, 2021. 2

[14] Jianzhong He, Xu Jia, Shuaijun Chen, and Jianzhuang Liu. Multi-source domain adaptation with collaborative learning for semantic segmentation. In *CVPR*, 2021. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[16] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *CVPR*, 2021. 1, 2

[17] Lei Huang, Dawei Yang, Bo Lang, and Jia Deng. Decorrelated batch normalization. In *CVPR*, 2018. 1

[18] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 3, 4

[19] Juwon Kang, Sohyun Lee, Namyup Kim, and Suha Kwak. Style neophile: Constantly seeking novel styles for domain generalization. In *CVPR*, 2022. 1, 2, 4

[20] Jin Kim, Jiyoung Lee, Jungin Park, Dongbo Min, and Kwanghoon Sohn. Pin the memory: Learning to generalize semantic segmentation. In *CVPR*, 2022. 1, 2, 5, 6, 7, 8

[21] Shu Kong and Charless C Fowlkes. Recurrent pixel embedding for instance grouping. In *CVPR*, 2018. 5

[22] Varun Ravi Kumar, Marvin Klingner, Senthil Yogamani, Stefan Milz, Tim Fingscheidt, and Patrick Mader. Syndistnet: Self-supervised monocular fisheye camera distance estimation synergized with semantic segmentation for autonomous driving. In *WACV*, 2021. 1

[23] Jean Lahoud, Bernard Ghanem, Marc Pollefeys, and Martin R Oswald. 3d instance segmentation via multi-task metric learning. In *ICCV*, 2019. 2, 5

[24] Seunghun Lee, Sunghyun Cho, and Sunghoon Im. Dranet: Disentangling representation and adaptation networks for unsupervised cross-domain adaptation. In *CVPR*, 2021. 2

[25] Suhyeon Lee, Hongje Seong, Seongwon Lee, and Euntai Kim. Wildnet: Learning domain generalized semantic segmentation from the wild. In *CVPR*, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[26] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018. 2, 5, 6, 7, 8

[27] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, 2017. 2

[28] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *ICCV*, 2019. 2

[29] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *CVPR*, 2018. 1, 2

[30] Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. Prototypical contrastive learning of unsupervised representations. In *ICLR*, 2020. 2

[31] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *ECCV*, 2018. 2

[32] Road Augmentation Library. `https://github.com/UjjwalSaxena/Automold--Road-Augmentation-Library`. 5

[33] Ping Luo, Jiamin Ren, Zhanglin Peng, Ruimao Zhang, and Jingyu Li. Differentiable learning-to-normalize via switchable normalization. In *ICLR*, 2018. 1

[34] Xinyu Luo, Jiaming Zhang, Kailun Yang, Alina Roitberg, Kunyu Peng, and Rainer Stiefelhagen. Towards robust semantic segmentation of accident scenes via multi-source mixed sampling and meta-learning. In *CVPR*, 2022. 2

[35] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *ECCV*, 2018. 5

[36] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *ICML*, 2013. 1

[37] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *CVPR*, 2018. 2

[38] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017. 5

[39] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *CVPR*, 2020. 2

[40] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *ECCV*, 2018. 1, 2, 3, 5, 6, 7, 8

[41] Xingang Pan, Xiaohang Zhan, Jianping Shi, Xiaoou Tang, and Ping Luo. Switchable whitening for deep representation learning. In *ICCV*, 2019. 1

[42] Duo Peng, Yinjie Lei, Munawar Hayat, Yulan Guo, and Wen Li. Semantic-aware domain generalized segmentation. In *CVPR*, 2022. 1, 2, 5, 6, 7, 8

[43] Duo Peng, Yinjie Lei, Lingqiao Liu, Pingping Zhang, and Jun Liu. Global and local texture randomization for synthetic-to-real semantic segmentation. *IEEE Transactions on Image Processing*, 30:6594–6608, 2021. 2

[44] Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. Efficient domain generalization via common-specific low-rank decomposition. In *ICML*, 2020. 2

[45] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *CVPR*, 2021. 1

[46] Tiago Ramalho and Marta Garnelo. Adaptive posterior learning: few-shot learning with a surprise-based memory module. In *ICLR*, 2018. 5

[47] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016. 5

[48] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016. 5

[49] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018. 5

[50] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 2

[51] Herbert A Simon and Allen Newell. Human problem solving: The state of the theory in 1970. *American Psychologist*, 26(2):145, 1971. 2

[52] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017. 2

[53] Haitao Tian, Shiru Qu, and Pierre Payeur. A prototypical knowledge oriented adaptation framework for semantic segmentation. *IEEE Transactions on Image Processing*, 31:149–163, 2021. 2

[54] Zhuotao Tian, Xin Lai, Li Jiang, Shu Liu, Michelle Shu, Hengshuang Zhao, and Jiaya Jia. Generalized few-shot semantic segmentation. In *CVPR*, 2022. 2

[55] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, 2018. 2

[56] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *CVPR*, 2017. 1

[57] SS Vallender. Calculation of the wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786, 1974. 4

[58] Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and CV Jawahar. Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *WACV*, 2019. 5

[59] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *NeurIPS*, 2018. 2

[60] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, 2019. 2

[61] Haoran Wang, Tong Shen, Wei Zhang, Ling-Yu Duan, and Tao Mei. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In *ECCV*, 2020. 2

[62] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 1

[63] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *ICCV*, 2019. 2

[64] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *ICCV*, 2021. 2, 5

[65] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 2

[66] Qi Xu, Liang Yao, Zhengkai Jiang, Guannan Jiang, Wenqing Chu, Wenhui Han, Wei Zhang, Chengjie Wang, and Ying Tai. Dirl: Domain-invariant representation learning for generalizable semantic segmentation. In *AAAI*, 2022. 1, 2

[67] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. In *NeurIPS*, 2020. 2

[68] Zhenlin Xu, Deyi Liu, Junlin Yang, Colin Raffel, and Marc Niethammer. Robust and generalizable visual representation learning via random convolutions. In *ICLR*, 2021. 2

[69] Yi Yang, Yueting Zhuang, and Yunhe Pan. Multiple knowledge representation for big data artificial intelligence: framework, applications, and case studies. *Frontiers of Information Technology & Electronic Engineering*, 22(12):1551–1558, 2021. 2

[70] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020. 5

[71] Yunlong Yu, Zhong Ji, Jungong Han, and Zhongfei Zhang. Episode-based prototype generating network for zero-shot learning. In *CVPR*, 2020. 2

[72] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *ICCV*, 2019. 1, 2

[73] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *CVPR*, 2021. 2

[74] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[75] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *ICLR*, 2020. 1, 2, 3

[76] Tianfei Zhou, Wenguan Wang, Ender Konukoglu, and Luc Van Gool. Rethinking semantic segmentation: A prototype view. In *CVPR*, 2022. 2, 4

[77] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, 2018. 2