Style Blind Domain Generalized Semantic Segmentation via Covariance Alignment and Semantic Consistence Contrastive Learning

Woo-Jin Ahn[1] Geun-Yeong Yang[1] Hyun-Duck Choi[2*] Myo-Taeg Lim[1*]
[1]Korea University [2]Chonnam National University
[1]{wjahn,hggofficial,mlim}@korea.ac.kr [2]ducky.choi@jnu.ac.kr

Abstract

Deep learning models for semantic segmentation often experience performance degradation when deployed to unseen target domains unidentified during the training phase. This is mainly due to variations in image texture (i.e. style) from different data sources. To tackle this challenge, existing domain generalized semantic segmentation (DGSS) methods attempt to remove style variations from the feature. However, these approaches struggle with the entanglement of style and content, which may lead to the unintentional removal of crucial content information, causing performance degradation. This study addresses this limitation by proposing BlindNet, a novel DGSS approach that blinds the style without external modules or datasets. The main idea behind our proposed approach is to alleviate the effect of style in the encoder whilst facilitating robust segmentation in the decoder. To achieve this, BlindNet comprises two key components: covariance alignment and semantic consistency contrastive learning. Specifically, the covariance alignment trains the encoder to uniformly recognize various styles and preserve the content information of the feature, rather than removing the style-sensitive factor. Meanwhile, semantic consistency contrastive learning enables the decoder to construct discriminative class embedding space and disentangles features that are vulnerable to misclassification. Through extensive experiments, our approach outperforms existing DGSS methods, exhibiting robustness and superior performance for semantic segmentation on unseen target domains. The code is available at https://github.com/root0yang/BlindNet.

1. Introduction

Semantic segmentation, a technique that classifies each pixel in an image into predefined categories, has garnered significant attention due to its potential applications in various fields. Particularly, it plays a crucial role in autonomous driving [1, 18] and robotic systems [34, 35]. Besides, with the advent of large datasets, deep neural networks have emerged as a trending approach for semantic segmentation tasks, achieving impressive results [3, 5, 44, 54]. However, there remains a major bottleneck detailing the meticulous and labor-intensive process of dataset labeling. More particularly, this process not only consumes time but also poses economic challenges [8, 46]. To address this challenge, synthetic datasets have emerged as a compelling alternative. Specifically, these datasets, generated using three-dimensional (3D) rendering techniques, offer vast amounts of easily accessible data, eliminating the need for manual labeling [43, 45]. However, a challenge arises when models trained on synthetic datasets are deployed in real-world scenarios. More precisely, a domain shift problem arises due to style factor discrepancies (*e.g.* texture, illumination, and image quality) between synthetic and real-world data, which affects the performance of the model, as shown in Fig. 1.

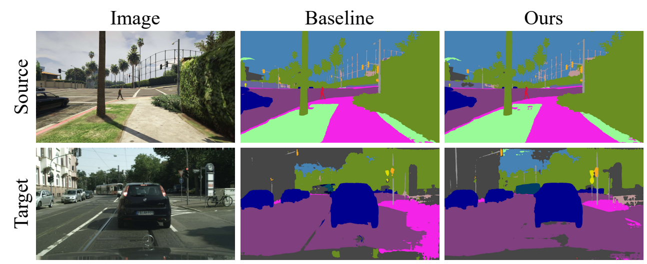


Figure 1. Comparison of semantic segmentation results between the baseline (DeepLabV3+ with ResNet50 backbone) and our BlindNet. Both models are trained on the source domain (GTAV [43]) and tested on the target domain (Cityscapes [8]).

To address the domain shift problem, domain adaptation semantic segmentation (DASS) has been introduced [15, 17, 22, 25, 51, 62]. Specifically, DASS aims to bridge the gap between source and target domains by aligning their data distributions. However, a significant limitation of DASS is its dependency on the target domain during train-

*Corresponding Authors

ing. For DASS to function effectively, target domain samples must be available during the training phase.

Meanwhile, domain generalized semantic segmentation (DGSS) has been introduced as an alternative approach to tackle the domain shift problem. Unlike DASS, DGSS is trained only with the source domain, aiming to extract domain-invariant features. To achieve this, two main techniques have been employed: domain randomization (DR) and feature normalization (FN).

DR augments the training set by introducing variability, either by altering the image style [41, 60] or by modifying the feature representation [24, 53]. By exposing the model to a wider variety of styles via DR, the network is less likely to overfit to the specific styles present in the training data. Consequently, the robustness of the model is improved, making it more adept at generalizing to new, unseen domains. Nevertheless, a crucial limitation of DR is its significant dependence on auxiliary domains.

FN methods, converse to DR, regularize the features to prevent the model from overfitting to the distinct styles or characteristics of the training data. This is achieved by removing domain-specific style information using feature statistics, such as instance normalization [38] or whitening transformation [7, 19, 39, 42]. While these approaches effectively remove style-related information, they simultaneously pose the challenge of removing semantic content because content and style information are entangled. Consequently, the model fails to capture the essential patterns or features required for accurate segmentation prediction.

To address this problem, we propose BlindNet, a model that blinds the style within the encoder and improves the robustness of the decoder, without requiring auxiliary datasets or external modules. Specifically, our proposed BlindNet consists of two components: covariance alignment for the encoder and semantic consistency contrastive learning for the decoder. Precisely, the covariance alignment facilitates the generation of style-invariant features with the proposed covariance matching loss function (CML) and the cross-covariance loss function (CCL). Specifically, CML mitigates the effects of style variations, while CCL focuses on preserving content information, effectively addressing the prevalent content information loss observed in the FN method. To further improve the generalization ability, we develop semantic consistency contrastive learning, which consists of class-wise contrastive learning (CWCL) and semantic disentanglement contrastive learning (SDCL). Particularly, the CWCL constructs a discriminative class embedding space, while SDCL disentangles features of similar classes that often lead to prediction errors. Extensive experiments across various datasets demonstrate that the proposed BlindNet outperforms existing DGSS methods.

Our contributions are summarized as follows:

- We propose a covariance alignment within the encoder, comprising CML and CCL. Specifically, the CML aims to mitigate the effects of style variations, while CCL ensures the preservation of content information, together facilitating the generation of style-agnostic features.
- We propose semantic consistency contrastive learning within the decoder that comprises CWCL and SDCL, utilizing segmentation masks. Specifically, CWCL generates discriminative embeddings, whereas SDCL disentangles features of similar classes, enhancing the robustness of the model.
- Through extensive experiments, we demonstrate the superiority of our approach in DGSS, without the need to alter the network architecture or rely on external datasets.

2. Related Works

Domain adaptation and generalization for semantic segmentation. Domain adaptation (DA) aims at minimizing the distribution discrepancy between different domains, enabling a model to generalize from a source to a target domain. For DASS, adversarial training and cross-domain self-training strategies are commonly used. Particularly, adversarial-based methods [15, 29] employ generative adversarial networks [11] to close the feature distribution gap between source and target domains. Meanwhile, cross-domain self-training methods [16, 37, 59, 62] generate pseudo-labels for target domain data using pre-trained models, and employ them as training data, thereby expanding the training data and reducing distribution differences between the source and target domain.

Domain generalization (DG) methods [26, 28, 30, 50, 61] aim to improve the generalization ability of the model without accessing the target domain during training. Since the difference in style of the image is the main cause of the disparity, most existing domain generalized semantic segmentation methods utilize the style information of the image for domain-invariant learning. Feature statistics (*e.g.* mean, variance, covariance, gram matrix, *etc.*) which are commonly used in style transfer [10, 21, 27, 57] are employed to capture the style information. Interestingly, existing DGSS methods can be separated into two parts: i) domain randomization to expand the distribution of style or ii) feature normalization to remove style.

DR involves randomizing either the image or its features through stylization to learn domain-invariant features from various styles. For example, Peng *et al.* [42] extended the source domain data by stylizing images in the style of unreal paintings. Similarly, Yue *et al.* [60] and Huang *et al.* [22] attempted to enhance generalization by synthesizing images with diverse styles in the image space. In another study, Lee *et al.* [24] adopted ImageNet data [9] as wild data and performed randomization via synthesis in the feature space. Meanwhile, Wu *et al.* [53] diversified the trainable feature space by mixing the statistics of the feature and its colorjittered feature with Ada-IN [21].

FN methods aim to remove domain-specific styles from features, extracting only domain-invariant content. For instance, Pan *et al*. [39] first attempted the DGSS method, combining batch normalization (BN) [23] and instance normalization (IN) [48] in the network layer. While BN preserves the content information within discriminative features, IN focuses on removing domain-specific style information from features. In a study, Choi *et al*. [7] addressed the limitations of previous whitening transformation [6, 31] that can eliminate the content information from the feature. Specifically, they proposed an instance-selective whitening approach designed to remove covariance components that are sensitive to domain shifts. Peng *et al*. [41] developed semantic-aware normalization that performs on class-wise and semantic-aware whitening that aligns channels based on the prediction through group whitening transformation [6]. In a study, Xu *et al*. [55] introduced the prior guided attention module and guided feature whitening to re-calibrate the feature and remove domain-specific style effects, respectively. Unlike the FN methods that directly remove the style component, our work explores a covariance alignment method that mitigates the effect of the style's effect while preserving the content information. Our method achieves the DGSS without any additional modules or auxiliary datasets.

Contrastive Learning. Contrastive learning aims to learn representations by maximizing the similarity between positive pairs of samples while minimizing the similarity between negative pairs. In recent years, it has attracted significant attention for its effectiveness in learning discriminative representations across various tasks [2, 4, 12, 14]. Oord *et al*. [36] were the first to introduce the infoNCE loss, a type of contrastive loss function designed for self-contrastive learning. In a work, Park *et al*. [40] introduced patch-level contrastive learning for image translation, using co-located patches as positive pairs and spatially distant patches as negatives to maintain image context. For the semantic segmentation task, Wang *et al*. [52] introduced class-wise contrastive learning to aid the model in learning the embedding space of each class. Specifically, they sampled the classes existing in the images and applied contrastive learning based on the class label. For DGSS, Lee *et al*. [24] adopted contrastive learning to learn the ImageNet information in their model. Specifically, they set the ImageNet data as wild and applied the contrastive learning method by setting the wild-stylized feature and its closest wild content as positive samples. In another study, Yang *et al*. [56] developed multi-level contrastive learning, which designed instance prototypes and class prototypes for contrastive learning. Specifically, they sample each class's pixel features and apply contrastive learning with a transition-probability ability matrix. Unlike recent DGSS works that embed the original image, we define contrastive learning for the augmented image. Specifically, the proposed method builds a robust embedding space by preserving the semantic consistency of the feature representation across various domains.

3. Method

The goal of the proposed method is to train a segmentation model φ on a given source domain S and generalize well to the unseen target domain T. Precisely, the source domain $S = \{(x, y)\}$ contains a paired image $x \in \mathbb{R}^{H\times W\times 3}$ and segmentation label $y \in \mathbb{R}^{H\times W\times C}$, where H, W, and C denote the height of the image, the width of the image, and the class number of the segmentation map, respectively. The model φ takes an original x and its augmented counterpart x_a, which have the same content but different styles, and uses the feature information to enhance its generalization ability.

As shown in Fig. 2, our method leverages the feature information through two main approaches: covariance alignment and semantic consistency contrastive learning. Specifically, the covariance matching ensures that features, having different styles but the same content, contain similar information. Additionally, the semantic consistency contrastive learning embeds the generalized features into discriminative representation based on the segmentation label.

3.1. Covariance Alignment

The domain shift in semantic segmentation results from changes in the visual characteristics of the image, known as style. This style information is typically detected in the shallow layers of networks [38], which are the encoders of the segmentation model. Based on this understanding, our method targets the encoder to handle style variations by employing the proposed covariance matching loss and cross-covariance loss function.

Covariance Matching Loss

To train the network to uniformly recognize various styles without removing content information, we introduce the CML. Specifically, the loss aims to minimize the difference between covariance matrices derived from different styles of image features. Given an image pair (x, x_a), the features from i^{th} block of the encoder are represented as $F^i \in \mathbb{R}^{(H^i\times W^i)\times C^i}$ and $F^i_a \in \mathbb{R}^{(H^i\times W^i)\times C^i}$, respectively. Further, following the methodologies of [7, 41], we compute the covariance matrices using instance normalized features [48], which ensures consistent scaling across features. The feature maps are normalized and flattened into $\bar{F}$ and $\bar{F_a} \in \mathbb{R}^{(HW)\times C}$, which are given by:

$$\bar{F} = \frac{(F - \mu(F))}{\sigma(F)}, \bar{F_a} = \frac{(F_a - \mu(F_a))}{\sigma(F_a)}, \tag{1}$$

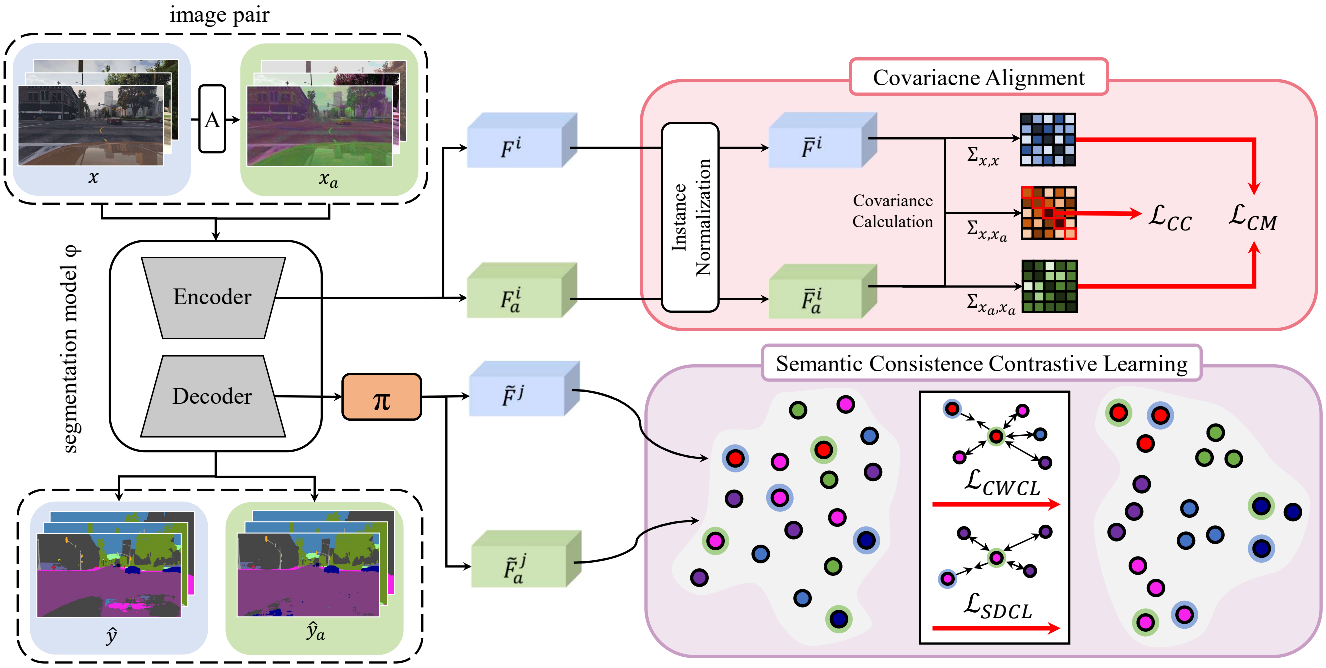


Figure 2. Overview of the proposed BlindNet. The network processes a pair of images - the original image x and its augmented counterpart x_a. It employs covariance alignment to treat encoder features and utilizes semantic consistency contrastive learning for the processing of decoder features.

where $\mu(\cdot) \in \mathbb{R}^C$ and $\sigma(\cdot) \in \mathbb{R}^C$ denote the mean and standard deviation of the features. Utilizing the normalized features, we evaluate the covariance matrices for the original and augmented image features as:

$$\Sigma^i_{x,x} = \bar{F^i}^T \cdot \bar{F^i}, \qquad \Sigma^i_{x_a,x_a} = \bar{F}^{i\,T}_a \cdot \bar{F}^i_a. \tag{2}$$

The CML is then formulated to align these covariance matrices, ensuring that the network maintains consistency in the presence of style variations. The CML is defined as:

$$\mathcal{L}_{CM} = \sum_{i=1}^{n_e} \|\Sigma^i_{x,x} - \Sigma^i_{x_a,x_a}\|_2, \tag{3}$$

where n_e denotes the number of blocks of the encoder.

Cross-covariance Loss

While the CML effectively aligns the internal distributions of features within the same image, it does not fully account for the direct correlations across paired images. To complement this, we introduce CCL, which aims to encode the consistent content information of an image pair (x, x_a) by utilizing the cross-covariance of the image pair. Given the normalized feature pair $(\bar{F^i}, \bar{F^i_a})$, the cross-covariance of the feature pair can be expressed as:

$$\Sigma^i_{x,x_a} = \bar{F^i}^T \cdot \bar{F}^i_a. \tag{4}$$

The cross-covariance is expected to exhibit an identity matrix, as the feature pair should contain identical information. Nonetheless, the proposed CCL converges only the diagonal component of the covariance matrix to one. This is to prevent the drawbacks of the existing FN methods [7, 42] from removing content information. The CCL function is thus defined as:

$$\mathcal{L}_{CC} = \sum_{i=1}^{n_e} \|\text{diag}(\Sigma^i_{x,x_a}) - \mathbb{1}\|_2, \tag{5}$$

where $\text{diag}(\Sigma_c) \in \mathbb{R}^C$ denotes the column vector comprising diagonal elements of Σ^i_{x,x_a} and $\mathbb{1} \in \mathbb{R}^C$ denotes the one vector.

3.2. Semantic Consistence Contrastive Learning

While the encoder focuses on generating style-blinded features, the decoder aims to improve the robustness of the segmentation prediction against domain shifts. For the decoder, we employ a contrastive learning approach, which has demonstrated effectiveness in extracting discriminative features [4]. Specifically, we utilize the InfoNCE loss [36], which is formulated as:

$$\mathcal{L}_{IN}(a, p, n) = -\log\left(\frac{e^{(a\cdot p/\tau)}}{e^{(a\cdot p/\tau)} + \sum_n^{N^-} e^{(a\cdot n/\tau)}}\right) \tag{6}$$

where a, p, n, and N^- denote anchor, positive sample, negative sample, and negative sample set, respectively.

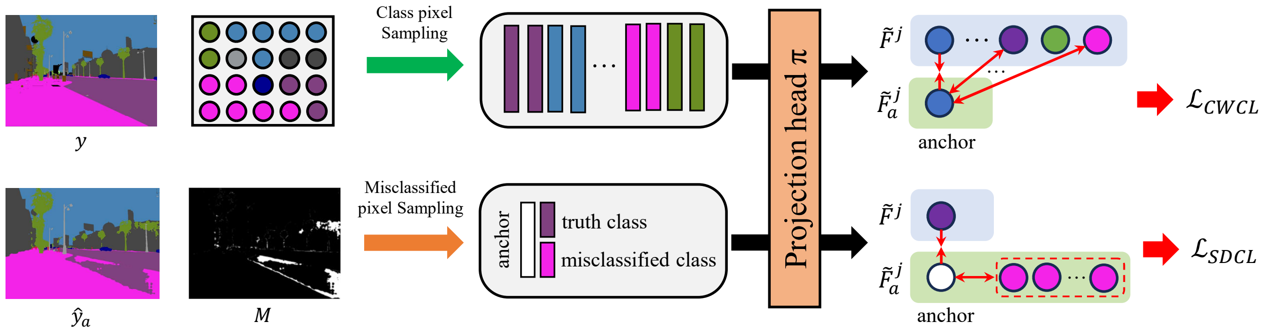


Figure 3. Illustration of semantic consistency contrastive learning: The mask M represents the error mask derived from the augmented segmentation map. CWCL conducts contrastive learning by sampling per segmentation class and SDCL conducts contrastive learning based on the M. Both methods share a projection head π for the semantic representation.

To achieve consistent feature representation in DGSS across various styles, we introduce semantic consistency contrastive learning. Specifically, the anchor is derived from the augmented image x_a, while the positive sample is extracted from the corresponding pixel of the original image x. Our method consists of two main components based on the negative sample as shown in Fig. 3: class-wise contrastive learning and semantic disentanglement contrastive learning.

Class-wise Contrastive Learning

Our CWCL aims to build a discriminative embedding space for each segmentation class using different classes of the original image as negatives. Given an image pair (x, x_a), the features from the j^{th} block of the decoder at pixel position (m, n) are denoted as $F^j_{(m,n)}$ and $F^j_{a,(m,n)} \in \mathbb{R}^{(1\times1)\times C^j}$, where C^j indicates the channel length of the feature. As mentioned above, we take $F^j_{a,(m,n)}$ as the anchor and $F^j_{(m,n)}$ as the positive sample, since they represent the same content at the corresponding spatial location. To obtain the negative samples from F^j, we leverage the resized segmentation class label $y^j \in \mathbb{R}^{(H^j\times W^j)\times C}$ to collect the different class samples. The samples are passed through the projection head, denoted as π, resulting in the projected features $\tilde{F}$ and $\tilde{F}_a$. We define our CWCL as:

$$\mathcal{L}_{CWCL} = \sum_{j}^{n_d} \mathcal{L}_{IN}\left(\tilde{F}^j_{a,(m,n)}, \tilde{F}^j_{(m,n)}, \tilde{F}^j_{(p,q)}\right) \quad (7)$$
$$\text{where} \quad (p,q) \in \{(p,q) \in P | y^j_{(p,q)} \neq y^j_{(m,n)}\}$$

The set P represents all pixel positions in feature F^j, with dimensions $H^j \times W^j$ corresponding to the height and width. Additionally, n_d denotes the total number of blocks in the decoder.

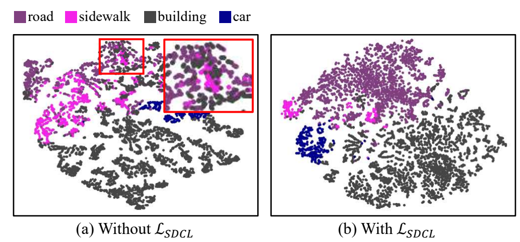


Figure 4. t-SNE [49] visualization comparing scenarios with and without $\mathcal{L}_{SCCL}$. In (b), the application of SCCL results in a clear separation between the sidewalk (pink), the road (purple), and the building (gray).

Semantic Disentanglement Contrastive Learning

Domain shifts can lead to the entanglement of similar classes, causing the model to misclassify, as illustrated in Fig. 4. To mitigate this issue, we introduce the SDCL, specifically designed to disentangle the feature x_a that has been misclassified, making it closer to the correct class and far from the misclassified class to achieve effective disentanglement. To further ensure consistent feature space and capture the semantic meaning, we share the projection head π used in the CWCL loss. Given the predicted segmentation map of the augmented image, represented as $\hat{y}_a = \varphi(x_a)$, we resize it to $\hat{y}^j_{a,(m,n)} \in \mathbb{R}^{(H^j\times W^j)\times C}$. Similarly, $y^j_{(m,n)}$ represents the ground truth segmentation map. Using these segmentation maps, we set the anchor at positions where $\hat{y}^j_{a,(m,n)} \neq y^j_{(m,n)}$. Negative samples are selected from the augmented image features corresponding to the anchor's misclassified class. The samples go through the projection head π. Our SDCL loss is defined as follows:

$$\mathcal{L}_{SDCL} = \sum_{j}^{n_d} \mathcal{L}_{IN}\left(\tilde{F}^j_{a,(m,n)}, \tilde{F}^j_{(m,n)}, \tilde{F}^j_{a,(r,s)}\right) \quad (8)$$
$$\text{where} \quad (r,s) \in \{(r,s) \in P | y^j_{(r,s)} = \hat{y}^j_{(m,n)}\}$$

Finally, combining the cross-entropy segmentation loss

Backbone	Methods	External Dataset	External Module	Trained on GTAV (G): C	B	M	S	Trained on Cityscapes (C): B	M	S	G
ResNet50 [13]	Baseline [3]	-	-	28.95	25.14	28.18	26.23	44.96	51.68	23.29	42.55
	IBN-Net [38]	-	-	33.85	32.30	37.75	27.90	48.56	57.04	26.14	45.06
	RobustNet [7]	-	-	37.31	35.20	40.33	28.30	50.73	58.64	26.20	45.00
	SiamDoGe [53]	-	-	42.96	37.54	40.64	28.34	51.53	59.00	26.67	45.08
	DIRL [55]	-	✓	41.04	39.15	41.60	-	51.80	-	26.50	46.52
	WildNet [24]	✓	-	44.62	38.42	46.09	31.34	50.94	58.79	27.95	47.01
	SANSAW [42]	-	✓	39.75	37.34	41.86	30.79	**52.95**	59.81	28.32	47.28
	SPC [20]	-	✓	44.10	40.46	45.51	-	-	-	-	-
	DPCL [56]	-	✓	44.74	40.59	46.33	30.81	50.97	58.59	25.85	46.00
	Ours	-	-	**45.72**	**41.32**	**47.08**	**31.39**	51.84	**60.18**	**28.51**	**47.97**
ShuffleNetV2 [32]	Baseline [3]	-	-	25.56	22.17	28.60	23.33	36.84	43.13	21.56	36.95
	IBN-Net [38]	-	-	27.10	31.82	34.89	25.56	41.89	46.35	22.99	40.91
	RobustNet [7]	-	-	30.98	32.06	35.31	24.31	41.94	46.97	22.82	40.17
	SiamDoGe [53]	-	-	34.40	34.23	35.87	21.95	42.61	47.48	23.13	40.93
	DIRL [55]	-	✓	31.88	32.57	36.12	-	42.55	-	**23.74**	**41.23**
	DPCL [56]	-	✓	36.66	34.35	39.92	22.66	43.90	48.95	22.47	41.07
	Ours	-	-	**38.56**	**34.51**	**40.11**	**25.64**	**44.22**	**49.69**	23.54	41.10

Table 1. Quantitative comparison of mIoU (%) between DGSS methods. External dataset denotes the necessity of an auxiliary dataset during training and External module denotes the requirement of an additional module during inference. G, C, B, M, and S denote GTAV, Cityscapes, BDD100K, Mapillary, and SYNTHIA, respectively. The best and second-best results are **bolded** and underlined, respectively.

$\mathcal{L}_{CE}$ with other loss components, the total is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{CE} + \omega_1 \mathcal{L}_{CM} + \omega_2 \mathcal{L}_{CC} + \omega_3 \mathcal{L}_{CWCL} + \omega_4 \mathcal{L}_{SDCL} \tag{9}$$

where ω_1, ω_2, ω_3, and ω_4 denote the weighting factor of each loss functions.

4. Experiment

In this section, we describe the implementation details, the experimental setup for comparison with existing DGSS methods, and the ablation study conducted to further validate the effectiveness of our approach.

4.1. Implementation Details

We adopt DeepLabV3+ [3] for the segmentation architecture and use ResNet-50 [13], ShuffleNetV2 [32], and MoblieNetV2 [47] as the backbone network of the segmentation network. The model is trained for 40K iterations with a batch size of 8 using the SGD optimizer, which has a momentum of 0.9 and a weight decay of 5e-4. We employ a polynomial learning rate schedule with an initial rate of 1e-2 and a power of 0.9. For the simulation of domain shift, we augment the image x_a using strong color jittering transformation similar to [7]. The weighting parameters of (9), ω_1, ω_2, ω_3 and ω_4, are set as 0.2, 0.2, 0.3, and 0.3 respectively.

4.2. Datasets

We use two synthetic datasets (GTA [43] and SYNTHIA [45]), and three real-world datasets (Cityscapes [8], BDD-100K [58], and Mapillary [33]) for the experiment. All segmentation labels are evaluated based on 19 object categories.

GTAV (G) [43] is a large-scale dataset generated from the Grand Theft Auto V (GTAV) game engine. It comprises 24,966 images, split into 12,403 for training, 6,382 for validation, and 6,181 for testing with a resolution of 1914×1052.

SYNTHIA (S) [45] is a virtual, photo-realistic urban scene dataset comprising 9,400 images with a resolution of 960×720. Among these, 2,820 images are designated for evaluation.

Cityscapes (C) [8] is a large-scale urban scene dataset captured from 50 cities, primarily in Germany. Particularly, it contains 5,000 high-resolution images with a resolution of 2048×1024. The dataset is divided into 2,975 images for training, 500 for validation, and 1,525 for testing.

BDD-100K (B) [58] is another real-world urban scene dataset that contains more diverse 10000 urban driving scene images with a resolution of 1280×720. Specifically, the validation split (1,000 images) is used for evaluation.

Mapillary (M) [33] contains 25,000 images with a minimum resolution of 1920×1080, collected from various locations worldwide. Specifically, the validation split of 2,000 images is used for evaluation.

4.3. Comparison with DGSS methods

We compare our methods with other state-of-the-art DGSS methods: Baseline (DeepLabV3+ [3] trained on the source domain), IBN-Net [38], RobustNet [7], SiamDoGe [53],

Methods	External Module	Trained on GTAV (G)			
		C	B	M	Mean
Baseline [3]		25.94	25.73	26.45	26.04
IBN-Net [38]		30.14	27.66	27.07	28.29
RobustNet [7]		30.86	30.05	30.67	30.52
SiamDoGe [53]		34.15	34.50	32.34	33.67
DIRL [55]	✓	34.67	32.78	34.31	33.92
DPCL [56]	✓	<u>37.57</u>	<u>35.45</u>	<u>40.30</u>	<u>37.77</u>
Ours		**37.66**	**36.10**	**40.40**	**38.05**

Table 2. Quantitative comparison of mIoU (%) using MobileNetV2 [47] backbone trained on the GTAV dataset.

Methods	Externel Module	Params (M)	GFLOPS	Time (ms)
Baseline [13]		45.08	277.77	10.01
SANSAW [42]	✓	25.63	421.86	68.96
SPC [20]	✓	45.22	286.09	12.24
DIRL [55]	✓	45.41	278.11	11.69
DPCL [56]	✓	56.46	1188.64	823.78
Ours		45.08	277.78	10.03

Table 3. Computational cost comparison conducted using DeepLabV3+ with a ResNet-50 backbone on an NVIDIA Tesla V100 GPU with an image resolution of 2048×1024. Inference time is averaged over 400 trials.

$\mathcal{L}_{CM}$	$\mathcal{L}_{CC}$	$\mathcal{L}_{CWCL}$	$\mathcal{L}_{SDCL}$	C	B	M
				28.95	25.14	28.18
✓				38.08	36.65	40.62
✓	✓			40.42	37.81	43.91
		✓		42.03	38.27	44.02
		✓	✓	43.16	<u>38.59</u>	<u>45.38</u>
✓	✓	✓		<u>43.17</u>	38.23	44.84
✓	✓	✓	✓	**45.72**	**41.32**	**47.08**

Table 4. Ablation study on proposed losses. The experiments were conducted using DeepLabV3+ with ResNet-50 backbone, trained on the GTAV dataset. The losses are detailed in $\mathcal{L}_{CM}$: (3), $\mathcal{L}_{CC}$: (5), $\mathcal{L}_{CWCL}$: (7), $\mathcal{L}_{SDCL}$: (8)

DIRL [55], WildNet [24], SANSAW [42], SPC [20], and DPCL [56]. To evaluate the generalization ability of the model on arbitrary unseen domains, we conduct the experiment on two scenarios: i) trained on GTAV, tested on Cityscapes, BDD-100K, and Mapillary, and ii) trained on Cityscapes, tested on BDD-100K, Mapillary, and SYNTHIA. The quantitative results are computed with mean intersection over union (mIoU). Additionally, we compared the method trained on the backbone of ResNet-50 [13], ShuffleNetV2 [32], and MoblieNet [47], pre-trained on ImageNet [9].

Quantitative and Qualitative Results

Table 1 summarizes the quantitative results. Our method outperforms all other methods when trained on GTAV, using ResNet-50 as the backbone. When compared with FN methods that remove domain-specific styles, we demonstrate that our approach minimizes the loss of content information. We also show that our method effectively shows generalization ability when trained on Cityscapes. We further evaluate our methods with different backbones, showing the wide applicability of our method. When trained with ShuffleNetV2, our method achieves the first or second-best performance among unseen target domains. Table 2 shows the results of our method trained on GTAV with MobileNetV2, demonstrating the superiority of our method.

For qualitative evaluation, we compare the visual result between DGSS methods and ours. As depicted in Fig. 5, our method demonstrates superior results compared to other approaches, particularly in its overall prediction accuracy. Notably, our proposed techniques enable distinct prediction of features such as on road and sidewalk, yielding clearer segmentation boundaries. Please refer to the supplementary material for more qualitative results.

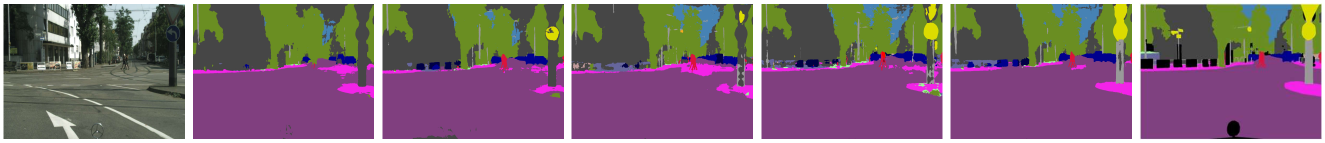
Computational cost analysis

To confirm that our approach does not incur additional computational overhead, we provide the number of parameters, GFLOPS, and average inference time of each method. As detailed in Table 3, our method operates comparably to baseline models by learning features intrinsically without adopting a separate module.

4.4. Ablation Studies

In this subsection, we conducted a series of ablation studies to demonstrate the individual contribution and effectiveness of each component within our method. Specifically, we investigate the impact of the following components: $\mathcal{L}_{CM}$, $\mathcal{L}_{CC}$, $\mathcal{L}_{CWCL}$, $\mathcal{L}_{SDCL}$. For the study, we use a scenario where the DeepLabV3+ model with backbone ResNet-50 model is trained on GTA and tested on Cityscapes, BDD-100K, and Mapillary.

Table 4 presents the impact of various proposed losses on domain generalization performance. Specifically, the baseline model, trained solely with cross-entropy loss, exhibits suboptimal performance on target domains because of overfitting to the source domain. Conversely, the integration of any proposed loss mechanisms leads to a marked enhancement in performance. More specifically, the incorporation of the covariance alignment ($\mathcal{L}_{CM}$, $\mathcal{L}_{CC}$) shows its efficacy in preserving essential content information by correlating features of paired images. The differential impact of the semantic consistency constrastive learning ($\mathcal{L}_{CWCL}$, $\mathcal{L}_{SDCL}$) is also evident, as it significantly aids in disentangling features of similar classes, thereby constructing a more robust embedding space.



Target Data | Baseline | RobustNet | WildNet | SPC | Ours | Ground Truth

Figure 5. Qualitative comparison between DGSS methods trained on GTAV (G) and tested on unseen target domains of Cityscapes (C) using DeeplabV3+ with ResNet50 backbone.

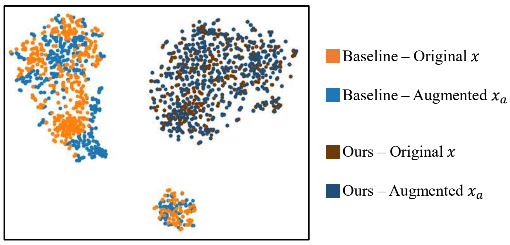


Figure 6. t-SNE [49] visualization comparing the covariance with and without $\mathcal{L}_{CML}$.

Cross-covariance loss			
Method	C	B	M
Whitening	38.68	36.91	42.12
Diagonal	40.42	37.81	43.91

(a)

# of classes			
#	C	B	M
10	45.57	38.88	46.37
15	45.72	41.32	47.08

(b)

# of negative samples			
#	C	B	M
10	44.76	38.21	46.46
50	45.72	41.32	47.08
100	44.44	39.14	46.29

(c)

Projection Head of SDCL			
MLP	C	B	M
Individual	44.91	38.45	46.31
Shared (SG)	44.03	38.15	46.64
Shared	45.72	41.32	47.08

(d)

Table 5. Ablation studies. (a) Calculation of $\mathcal{L}_{CC}$. (b) Number of classes for $\mathcal{L}_{CWCL}$. (c) Number of negative samples for $\mathcal{L}_{CWCL}$. (d) Projection head of $\mathcal{L}_{SDCL}$. "SG" indicates stop gradient.

Covariance Matching Loss. Fig. 6 presents t-SNE plots of covariances for original and augmented images, before and after the application of CML. The baseline network perceives original and augmented images differently from a style perspective. However, after applying CML, the distribution becomes more intermixed, indicating that our proposed CML effectively ensures similar recognition of different style images.

Calcuation of CCL. Table 5a demonstrates that our proposed cross-covariance method, which converges the diagonal components to 1, yields superior performance. As mentioned before, removing non-diagonal components, which contain content information actually degrades performance.

Sampling number in CWCL. Table 5b and Table 5c show the impact of varying the number of classes sampled per image and the number of samples per class in CWCL, respectively. As shown in Table 5b, the performance improves with an increase in the diversity of classes sampled in CWCL. This suggests that contrasting a broader array of classes enhances the model's discriminative capability. Furthermore, Table 5c demonstrates that a balanced number of negative samples per class leads to optimal performance.

Projection Head for SDCL. The influence of different project head configurations on the SDCL is investigated. We experimented with three distinct approaches: individual projection head, copying the weights of CWCL's (stop gradient), and shared projection head of CWCL. As demonstrated in Table 5d, sharing the projection head yielded the most superior results. The results indicate that SDCL not only relies on the semantic information from CWCL for effective disentanglement of similar classes but also enhances the embedding space learned by CWCL.

5. Conclusion

In this paper, we propose a novel BlindNet with covariance alignment and semantic consistency contrastive learning. By introducing covariance alignment, our method effectively addresses style variations, ensuring the extraction of features that are consistent across different styles. Furthermore, with the proposed semantic consistency contrastive learning, we not only facilitate the extraction of discriminative features but also enhance the generalization capabilities of the model in semantic segmentation predictions. Comprehensive experimental results validate the effectiveness of our approach, demonstrating its ability to generalize across multiple unseen target domains without requiring auxiliary domains or additional modules. Our future work will be improving and stabilizing the covariance alignment method.

Acknowledgement. This work was supported in part by the Basic Science Research Program through National Research Foundation of Korea (NRF) (Grants No. NRF-2022R1F1A1073543), the MSIT(Ministry of Science and ICT), Korea, under the ICAN(ICT Challenge and Advanced Network of HRD) support program(RS-2022-00156385) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation), and Innovative Human Resource Development for Local Intellectualization program through the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT)(IITP-2024-00156287).

References

[1] Florent Bartoccioni, Éloi Zablocki, Andrei Bursuc, Patrick Pérez, Matthieu Cord, and Karteek Alahari. Lara: Latents and rays for multi-camera bird's-eye-view semantic segmentation. In *Conference on Robot Learning*, pages 1663–1672. PMLR, 2023. 1

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 3

[3] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1, 6, 7

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3, 4

[5] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 1

[6] Wonwoong Cho, Sungha Choi, David Keetae Park, Inkyu Shin, and Jaegul Choo. Image-to-image translation via group-wise deep whitening-and-coloring transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10639–10647, 2019. 3

[7] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11580–11590, 2021. 2, 3, 4, 6, 7

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

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 7

[10] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 2

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[12] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 3

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 7

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[15] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr, 2018. 1, 2

[16] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9924–9935, 2022. 2

[17] Lukas Hoyer, Dengxin Dai, Qin Wang, Yuhua Chen, and Luc Van Gool. Improving semi-supervised and domain-adaptive semantic segmentation with self-supervised depth estimation. *International Journal of Computer Vision*, pages 1–27, 2023. 1

[18] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023. 1

[19] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6891–6902, 2021. 2

[20] Wei Huang, Chang Chen, Yong Li, Jiacheng Li, Cheng Li, Fenglong Song, Youliang Yan, and Zhiwei Xiong. Style projected clustering for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3061–3071, 2023. 6, 7

[21] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 2, 3

[22] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018. 1, 2

[23] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015. 3

[24] Suhyeon Lee, Hongje Seong, Seongwon Lee, and Euntai Kim. Wildnet: Learning domain generalized semantic segmentation from the wild. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pages 9936–9946, 2022. 2, 3, 6, 7

[25] Guangrui Li, Guoliang Kang, Wu Liu, Yunchao Wei, and Yi Yang. Content-consistent matching for domain adaptive semantic segmentation. In *European conference on computer vision*, pages 440–456. Springer, 2020. 1

[26] Lei Li, Ke Gao, Juan Cao, Ziyao Huang, Yepeng Weng, Xiaoyue Mi, Zhengze Yu, Xiaoya Li, and Boyang Xia. Progressive domain expansion network for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 224–233, 2021. 2

[27] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *Advances in neural information processing systems*, 30, 2017. 2

[28] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 624–639, 2018. 2

[29] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6936–6945, 2019. 2

[30] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 2

[31] Ping Luo. Learning deep architectures via generalized whitened neural networks. In *International Conference on Machine Learning*, pages 2238–2246. PMLR, 2017. 3

[32] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018. 6, 7

[33] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017. 6

[34] David Nilsson, Aleksis Pirinen, Erik Gärtner, and Cristian Sminchisescu. Embodied visual active learning for semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2373–2383, 2021. 1

[35] Yuya Onozuka, Ryosuke Matsumi, and Motoki Shino. Autonomous mobile robot navigation independent of road boundary using driving recommendation map. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4501–4508. IEEE, 2021. 1

[36] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3, 4

[37] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3764–3773, 2020. 2

[38] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 464–479, 2018. 2, 3, 6, 7

[39] Xingang Pan, Xiaohang Zhan, Jianping Shi, Xiaoou Tang, and Ping Luo. Switchable whitening for deep representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1863–1871, 2019. 2, 3

[40] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 319–345. Springer, 2020. 3

[41] 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, 3

[42] Duo Peng, Yinjie Lei, Munawar Hayat, Yulan Guo, and Wen Li. Semantic-aware domain generalized segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2594–2605, 2022. 2, 4, 6, 7

[43] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 102–118. Springer, 2016. 1, 6

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 1

[45] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 1, 6

[46] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10765–10775, 2021. 1

[47] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 6, 7

[48] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 3

[49] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 5, 8

[50] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31, 2018. 2

[51] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2517–2526, 2019. 1

[52] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7303–7313, 2021. 3

[53] Zhenyao Wu, Xinyi Wu, Xiaoping Zhang, Lili Ju, and Song Wang. Siamdoge: Domain generalizable semantic segmentation using siamese network. In *European Conference on Computer Vision*, pages 603–620. Springer, 2022. 2, 6, 7

[54] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 1

[55] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2884–2892, 2022. 3, 6, 7

[56] Liwei Yang, Xiang Gu, and Jian Sun. Generalized semantic segmentation by self-supervised source domain projection and multi-level contrastive learning. *arXiv preprint arXiv:2303.01906*, 2023. 3, 6, 7

[57] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9036–9045, 2019. 2

[58] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 6

[59] Fei Yu, Mo Zhang, Hexin Dong, Sheng Hu, Bin Dong, and Li Zhang. Dast: Unsupervised domain adaptation in semantic segmentation based on discriminator attention and self-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10754–10762, 2021. 2

[60] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2100–2110, 2019. 2

[61] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *International Conference on Learning Representations*, 2020. 2

[62] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. 1, 2