Semantic-Aware Domain Generalized Segmentation

Duo Peng[1] Yinjie Lei[1,*] Munawar Hayat[2] Yulan Guo[3] Wen Li[4]
[1]Sichuan University [2]Monash University [3]Sun Yat-sen University
[4]University of Electronic Science and Technology of China
duo_peng@stu.scu.edu.cn, yinjie@scu.edu.cn, munawar.hayat@monash.edu
guoyulan@sysu.edu.cn, liwenbnu@gmail.com

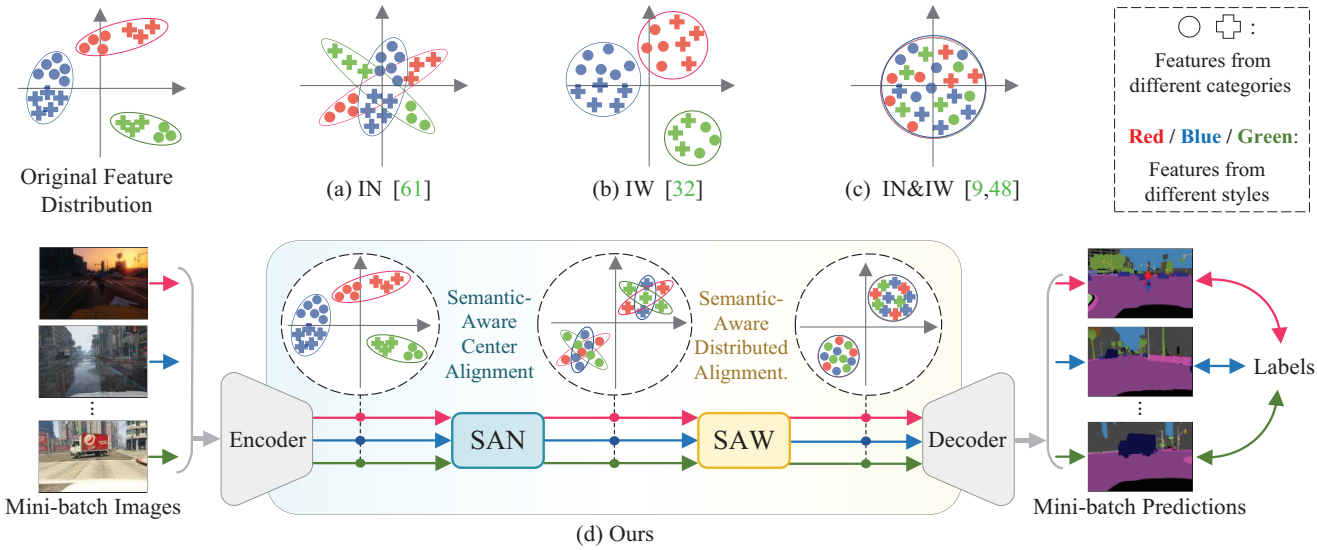


Figure 1. Illustration of existing Instance Normalization and Whitening methods and our proposed approach. (a-c) Existing methods broadly eliminate the global distribution variance but ignore the category-level semantic consistency resulting in limited feature discrimination. (d) Our proposed modules (SAN & SAW) encourage both intra-category compactness and inter-category separation through category-level feature alignment leading to both effective style elimination and powerful feature discrimination.

Abstract

*Deep models trained on source domain lack generalization when evaluated on unseen target domains with different data distributions. The problem becomes even more pronounced when we have no access to target domain samples for adaptation. In this paper, we address domain generalized semantic segmentation, where a segmentation model is trained to be domain-invariant without using any target domain data. Existing approaches to tackle this problem standardize data into a unified distribution. We argue that while such a standardization promotes global normalization, the resulting features are not discriminative enough to get clear segmentation boundaries. To enhance separation between categories while simultaneously promoting domain invariance, we propose a framework including two novel modules: **S**emantic-**A**ware **N**ormalization (**SAN**) and **S**emantic-**A**ware **W**hitening (**SAW**). Specifically, SAN focuses on category-level center alignment between features from different image styles, while SAW enforces distributed alignment for the already center-aligned features. With the help of SAN and SAW, we encourage both intra-category compactness and inter-category separability. We validate our approach through extensive experiments on widely-used datasets (*i.e. *GTAV, SYNTHIA, Cityscapes, Mapillary and BDDS). Our approach shows significant improvements over existing state-of-the-art on various backbone networks. Code is available at https://github.com/leolyj/SAN-SAW*

1. Introduction

Semantic segmentation is a critical machine vision task with multiple downstream applications, such as robotic nav-

*Corresponding Author: Yinjie Lei (yinjie@scu.edu.cn)

igation [25, 36, 41, 64], autonomous vehicles [17, 27, 50, 63] and scene parsing [28, 68, 69, 71]. While the current fully supervised deep learning based segmentation methods can achieve promising results when they are trained and evaluated on data from same domains [1, 4–6, 21, 37, 40, 67, 70], their performance dramatically degrades when they are evaluated on unseen out-of-domain data. To enable generalization of models across domains, different domain adaptation techniques have been recently proposed [3, 15, 16, 23, 44, 46, 55, 62, 75]. However, a critical limitation of domain adaptation methods is their reliance on the availability of target domain in advance for training purposes. This is impractical for many real-world applications, where it is hard to acquire data for rarely occurring concepts.

In this paper, we consider the challenging case of ***D**omain **G**eneralized **S**emantic **S**egmentation (**DGSS**)*, where we do not have access to any target domain data at the training time [9, 47, 48, 51, 66]. Existing methods tackle DGSS using two main approaches: (1) *Domain Randomization* [51, 66] which aims to increase the variety of training data by augmenting the source images to multiple domain styles. However, this is limiting since the augmentation schemes used are unable to cover different scenarios that may occur in the target domain. (2) *Normalization and Whitening* [9, 47, 48] which utilizes predefined Instance Normalization (IN) [61] or Instance Whitening (IW) [32] to standardize the feature distribution of different samples. IN separately standardizes features across each channel of individual images to alleviate the feature mismatch caused by style variations. However, as shown in Fig. 1 (a), IN only achieves center-level alignment and ignores the joint distribution among different channels. IW can remove linear correlation between channels, leading to well-clustered features of uniform distributions (see Fig. 1 (b)). Recent studies [9, 48] propose to combine IN and IW to achieve joint distributed feature alignment (see Fig. 1 (c)). Nevertheless, such global alignment strategy lacks the consideration of local feature distribution consistency. The features belonging to different object categories, which are originally well separated, are mapped together after normalization, leading to confusion among categories especially when generalizing to unseen target domains. Such semantic inconsistency inevitably results in sub-optimal training, causing performance degradation on unseen target domain and even the training domain (*i.e.* source domain).

To address the inherent limitations of IN and IW, we propose two modules, Semantic-Aware Normalization (SAN) and Semantic-Aware Whitening (SAW), which collaboratively align category-level distributions aiming to enhance the discriminative strength of features (see Fig. 1 (d)). Compared with traditional IN&IW based methods, our approach brings two appealing benefits: *First*, it carefully integrates semantic-aware center alignment and distributed alignment, enabling both discriminative and compact matching of features from different styles. Therefore, our method can significantly enhance models' generalization to out-of-domain distributed data. *Second*, existing methods improve the generalization ability at the cost of source domain performance [9, 47]. Nevertheless, our approach enhances the semantic consistency while improving category-level discrimination, thus leading to effective generalization with negligible performance drop on source domain.

Our extensive empirical evaluations on benchmark datasets show that our approach improves upon previous DGSS methods, setting new state-of-the-art performances. Remarkably, our method also performs favorably compared with existing SOTA domain adaptation methods that are trained using target domain data. In summary, followings are the major contributions of our work.

- We propose effective feature alignment strategies to tackle out-of-domain generalization for segmentation, without access to target domain data for training.
- The proposed semantic-aware alignment modules, SAN and SAW, are plug-and-play and can easily be integrated with different backbone architectures, consistently improving their generalization performance.
- Through extensive empirical evaluations, and careful ablation analysis, we show the efficacy of our approach across different domains and backbones, where it significantly outperforms the current state-of-the-art. Remarkably, we even perform at par with approaches using target domain data for training purposes.

2. Background

Here, we discuss recent approaches developed for Domain Adaptation (DA) and Domain Generalization (DG) in the context of semantic segmentation.

2.1. Domain Adaptation (DA)

Domain Adaptation seeks to narrow the domain gap between the source and target domain data. It aims to enhance the generalization ability of the model by aligning the feature distributions between the source and target images [15, 16, 38, 39, 58–60]. Domain adaptation for semantic segmenation (DASS) was first studied in [24, 72], and since then has gained significant research attention. We can broadly categorize the existing approaches for DASS into Adversarial Training, and Self-Training based methods. Most of the existing work on DASS has been dominated by *Adversarial Training* based approaches [7, 15, 23, 56, 73]. Inspired by Generative Adversarial Networks [20], these approaches are generative in nature, and synthesize indistinguishable features which are domain-invariant and deceive the domain classifier. *Self-Training* based DASS approaches are relevant once labeled training data is scarce.

These methods [35, 75] train the model with pseudo-labels which are generated from the previous models predictions. However, DA methods require access to the samples from the target domain, which limits their applicability on totally unseen target domain.

2.2. Domain Generalization (DG)

In contrast to Domain Adaptation, where the images in the target domain, although without labels, are accessible during the training process, Domain Generalization is evaluated on data from totally unseen domains [14, 43]. Domain generalization has been mostly explored on the image classification task, and a number of approaches have been proposed using as meta-learning [2, 29, 30, 34], adversarial training [31, 33, 52], autoencoders [18, 33], metric learning [12, 42] and data augmentation [19, 74]. The research on domain generalization for semantic segmentation (DGSS) is still in its infancy, with only a few existing approaches [9, 47, 48, 51, 66]. These existing DGSS methods mainly focus on two aspects: (1) Domain Randomization and (2) Normalization and Whitening. *Domain Randomization* based methods seek to synthesize images with different styles *e.g.* [66] leverages the advanced image-to-image translation to transfer a source domain image to multiple styles aiming to learn a model with high generalizability. Similarly, GTR [51] randomizes the synthetic images with the styles of unreal paintings in order to learn domain-invariant representations. *Normalization and Whitening Methods* apply different normalization techniques such as Instance Normalization (IN) [61] or whitening [48]. For example, based on the observation that Instance Normalization (IN) [61] prevents overfitting on domain-specific style of training data, [47] proposes to utilize IN to capture style-invariant information from appearance changes while preserving content related information. Inspired from [47], [47] proposes Switchable Whitening (SW), which combines IN with other whitening methods, aiming to achieve a flexible and generic features. In another recent approach [9], an instance selective whitening to disentangle domain-specific and domain-invariant properties is explored and only domain-specific features are normalized and whitened. However, all aforementioned methods perform a global alignment for features belong to different image categories. We aim to address this crucial limitation and propose an approach which enforces local semantic consistency during the trend of global style elimination.

3. Preliminaries

Let's denote an intermediate mini-batch feature map by $\mathbf{F} \in \mathbb{R}^{N\times K\times H\times W}$, where N, K, H and W are the dimensions of the feature map, *i.e. batch sample*, *channel*, *height* and *width*, respectively. $\mathbf{F}_{n,k,h,w} \in \mathbf{F}$ represent the feature element, where n, k, h, w respectively indicate the index of each dimension. Similarly, $\mathbf{F}_n \in \mathbb{R}^{K\times H\times W}$ denotes the features of n-th sample from mini-batch, and $\mathbf{F}_{n,k} \in \mathbb{R}^{H\times W}$ denotes the k-th channel of n-th sample.

Below, we first define Instance Normalization (IN) and Instance Whitening (IW), which have been commonly used by the existing approaches.

Instance Normalization (IN) simply standardizes features using statistics (*i.e.* mean and standard deviation) computed over each individual channel from each individual sample, given by:

$$\mathrm{IN}(\mathbf{F}) = \frac{\mathbf{F}_{n,k} - \mu_{n,k}}{\sigma_{n,k} + \varepsilon}, \tag{1}$$

where $\mathrm{IN}(\cdot)$ denotes the instance normalization process and ε is a small value to avoid division by zero. The mean $\mu_{n,k}$ and standard deviation $\sigma_{n,k}$ of n-th sample k-th channel are computed as follows:

$$\mu_{n,k} = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}\mathbf{F}_{n,k,h,w}, \tag{2}$$

$$\sigma_{n,k} = \sqrt{\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}(\mathbf{F}_{n,k,h,w} - \mu_{n,k})^2}. \tag{3}$$

Using above operations, IN transforms the features from different image samples to have a standard distribution, i.e, zero mean and one standard deviation. However, even though the features of each channel are centered and scaled into standard distribution, the joint distribution between channels might be mismatched.

Instance Whitening (IW) standardizes features by decorrelating the channels. As shown in Fig. 3 (a), it removes correlation between channels by making the covariance matrix close to the identity matrix through the following objective function:

$$\mathcal{L}_{\mathrm{IW}} = \sum_{n=1}^{N}||\Psi(\mathbf{F}_n) - \mathbf{I}||_1, \tag{4}$$

where $\Psi(\cdot)$ and $\mathbf{I}$ denotes the channel correlation and identity matrix. $\Psi(\mathbf{F}_n)$ is defined as:

$$\Psi(\mathbf{F}_n) = \begin{bmatrix} \mathrm{Cov}(\mathbf{F}_{n,1},\mathbf{F}_{n,1}) & \cdots & \mathrm{Cov}(\mathbf{F}_{n,1},\mathbf{F}_{n,K}) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(\mathbf{F}_{n,K},\mathbf{F}_{n,1}) & \cdots & \mathrm{Cov}(\mathbf{F}_{n,K},\mathbf{F}_{n,K}) \end{bmatrix}, \tag{5}$$

where $\mathrm{Cov}(\mathbf{F}_{n,i},\mathbf{F}_{n,j})$ is the covariance value between the i-th channel and j-th channel of feature $\mathbf{F}_n$, given by:

$$\mathrm{Cov}(\mathbf{F}_{n,i},\mathbf{F}_{n,j}) = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}(\mathbf{F}_{n,i,h,w}-\mu_{n,i})(\mathbf{F}_{n,j,h,w}-\mu_{n,j}). \tag{6}$$

IW [32] is capable of unifying the joint distribution shape through channel decorrelation for each sample. Combined with IN [61], IW can make a unified joint distributed

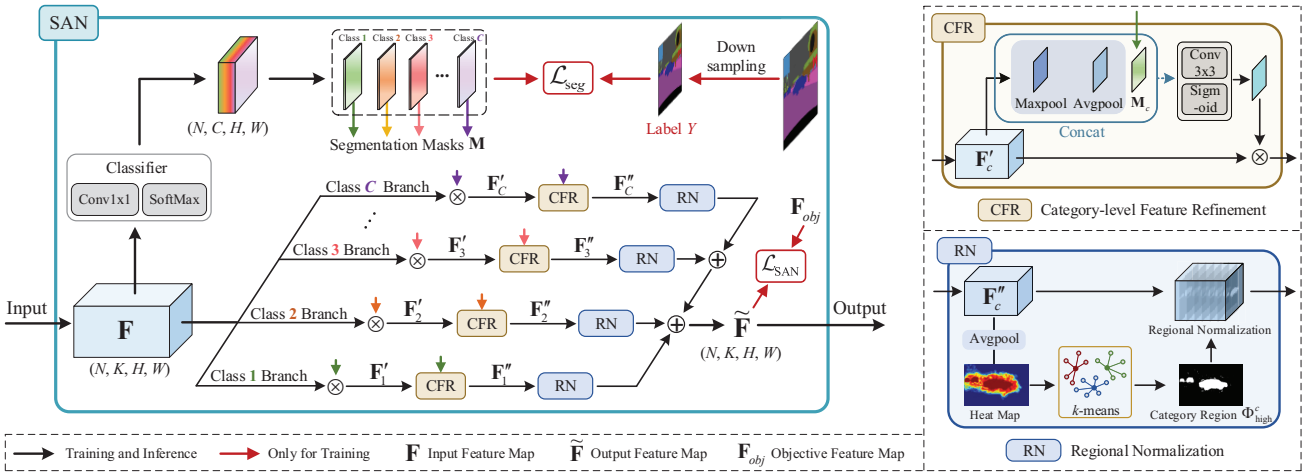


Figure 2. The detailed architecture of our Semantic Aware Normalization (SAN) module. SAN adapts a multi-branch normalization strategy, aiming to transform the feature map $\mathbf{F}$ into the category-level normalized features $\widetilde{\mathbf{F}}$, that are semantic-aware center aligned.

alignment. However, such global matching might cause some features to be mapped to an incorrect semantic category, resulting in poor segmentation boundary decisions. In the following Sec. 4, we introduce our method which aims to tackle these problems, and preserve the semantic relationships between different categories.

4. Proposed Method

Our goal is to achieve **semantic-aware center alignment** and **distributed alignment**. For this, we introduce two novel modules, ***Semantic-Aware Normalization (SAN)*** and ***Semantic-Aware Whitening (SAW)***. We sequentially embed these two modules in our network as shown in Fig. 1. We discuss these modules in detail below.

4.1. Semantic-Aware Normalization (SAN)

Given an intermediate mini-batch feature map $\mathbf{F}$, SAN transforms $\mathbf{F}$ into a feature map that is category-level centred. With the help of segmentation labels Y, we can easily obtain the desired objective feature map $\mathbf{F}_{obj}$ as:

$$\mathbf{F}_{obj} = \frac{\mathbf{F}^c_{n,k} - \mu^c_{n,k}}{\sigma^c_{n,k} + \varepsilon} \cdot \gamma^c + \beta^c, \tag{7}$$

$$\mu^c_{n,k} = \frac{1}{|Y(c)|} \sum_{Y(c)} \mathbf{F}^c_{n,k}, \tag{8}$$

$$\sigma^c_{n,k} = \sqrt{\frac{1}{|Y(c)|} \sum_{Y(c)} (\mathbf{F}^c_{n,k} - \mu^c_{n,k})^2}, \tag{9}$$

where $\mu^c_{n,k}$ and $\sigma^c_{n,k}$ are the mean and standard deviation computed from c-th category features of k-th channel, n-th sample, and the c-th category label $Y(c)$. Features $\mathbf{F}^c_{n,k}$ belong to c-th category in channel $\mathbf{F}_{nk}$. The weights for scaling and shifting are denoted by γ and β, respectively, which are both learnable parameters. We seperately allocate these affine parameters for each category (*i.e.* γ^c and β^c) aiming to adjust the standardized features from different categories to distinct spaces, thus making our feature space more discriminative. Note that different samples in the mini-batch share the same affine parameters in order to cast features of same category into a same feature space, thus ensuring category-level center alignment.

We utilize SAN to approximate Eq. (7) since the label Y is unavailable when testing on target domains. We follow a series of steps to transform the input features $\mathbf{F}$ into the objective features $\mathbf{F}_{obj}$ as shown in Fig. 2. First, we leverage the segmentation masks generated from the classifier to highlight the category region while simultaneously suppressing other regions in each normalization branch.

$$\mathbf{F}'_c = \mathbf{F} \otimes \mathbf{M}_c, \tag{10}$$

where $\mathbf{M}_c$ denotes the mask of the c-th category, $\otimes$ denotes Hadamard product and $\mathbf{F}'_c$ represents the masked feature map in c-th category branch. The generated mask can be too rough to precisely locate category features. We therefore introduce a Category-level Feature Refinement (CFR) block to adaptively adjust the highlighted features $\mathbf{F}'_c$ into $\mathbf{F}''_c$, given by:

$$\mathbf{F}''_c = \mathrm{Sigm}(f^{3\times3}([\mathbf{F}'_{c,max}; \mathbf{F}'_{c,avg}; \mathbf{M}_c])) \otimes \mathbf{F}'_c, \tag{11}$$

where $f^{3\times3}(\cdot)$ and $\mathrm{Sigm}(\cdot)$ denote 3×3 convolution and sigmoid function, respectively. $\mathbf{F}'_{c,max}$ and $\mathbf{F}'_{c,avg}$ are max-pooled and average-pooled features of feature map $\mathbf{F}'_c$. $[\mathbf{a}; \mathbf{b}]$ is the concatenation of $\mathbf{a}$ and $\mathbf{b}$ along channel axis.

In order to further refine the category-level center alignment, we design a Regional Normalization layer which normalizes features only within the category region instead of whole scene. After refinement, only the feature elements

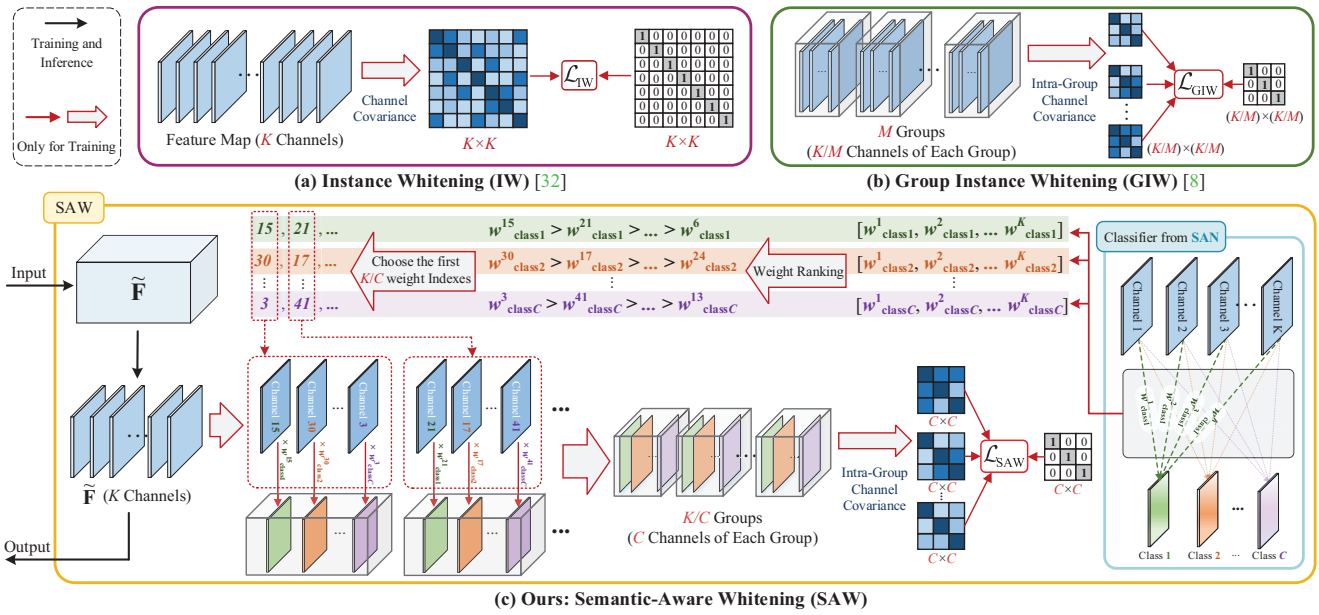


Figure 3. Illustration of feature whitening in IW [32], GIW [8] and the proposed SAW. **(a)** IW de-correlates all channels from each other. **(b)** GIW only de-correlates the channels in the same group. **(c)** SAW allocates channels related to different categories in each group.

with high value are assigned the category region. To flexibly identify feature elements, we apply k-means clustering on the spatial feature map obtained by avgpooling along channel axis. After dividing the spatial elements into k clusters, the clusters from the first to the t-th are considered to be the category region, and the remaining clusters are considered as ignored region. We set t to one and search the optimal k through the hyper-parameter search. In this paper, k is set to 5. See Sec. 5.5 for more details.

Thus, we can assign the feature elements of c-th branch into 2 clusters $\{\Phi^c_{\text{low}}, \Phi^c_{\text{high}}\}$, where Φ^c_{high} denotes the identified category region. We normalize features within the category region Φ^c_{high} for each individual channel. Finally, all category branches are added together, then re-shifted and scaled by the learnable affine parameters:

$$\widetilde{\mathbf{F}} = \sum_{c=1}^{C} \text{RN}(\mathbf{F}''_c, \Phi^c_{\text{high}}) \cdot \gamma^c + \beta^c, \tag{12}$$

where $\text{RN}(\cdot, \Phi^c_{\text{high}})$ denotes our regional normalization in c-th branch, γ^c and β^c are affine parameters as per Eq. (7). In order to ensure the processed feature map $\widetilde{\mathbf{F}}$ are category-level-center-aligned as per Eq. (7), we optimize the following cross-entropy loss:

$$\mathcal{L}_{\text{SAN}} = \text{CE}(\mathbf{M}, Y) + ||\widetilde{\mathbf{F}} - \mathbf{F}_{obj}||_1, \tag{13}$$

where $\mathbf{M} \in \{\mathbf{M}_1, \mathbf{M}_2, ..., \mathbf{M}_C\}$ denotes the set of predicted segmentation masks.

4.2. Semantic-Aware Whitening (SAW)

We propose the Semantic-Aware Whitening (SAW) module, to further enhance channel decorrelation for the distributed alignment of the already semantic-centred features. Instance Whitening (IW) is capable of unifying the joint distribution, which is useful for distributed alignment. However, directly adopting IW is not feasible, since such strong whitening that strictly removes correlation between all channels may damage the semantic content, resulting in loss of crucial domain-invariant information. Group Instance Whitening (GIW [8], shown in Fig. 3 (b)) is a simple solution to this problem. Given the feature map $\widetilde{\mathbf{F}}$ which is the output of SAN module, GIW is defined by:

$$\mathbf{G}^m_n = [\widetilde{\mathbf{F}}_{n,\frac{K(m-1)}{M}+1}; \widetilde{\mathbf{F}}_{n,\frac{K(m-1)}{M}+2}; \ldots; \widetilde{\mathbf{F}}_{n,\frac{Km}{M}}], \tag{14}$$

$$\mathcal{L}_{\text{GIW}} = \frac{1}{N}\sum_{n=1}^{N}\sum_{m=1}^{M} ||\Psi(\mathbf{G}^m_n) - \mathbf{I}||_1, \tag{15}$$

where $\mathbf{G}^m_n$ denotes the m-th group of n-th sample, M is the number of groups, and $\Psi(\cdot)$ is the channel correlation matrix defined in Eq. (4). While GIW improves the generalization by partial (group) channel decorrelation, it strictly decorrelates neighboring channels, lacking the consideration of searching more appropriate channel combinations. It is well-known that the channels in convolution neural networks are highly related to semantics. To this end, we propose SAW module to rearrange channels, ensuring each group contains channels related to different categories. Compared with grouping same-category-related channels, decorrelation between channels from different categories is more reasonable, since different-category-related channels activate different regions. Removing the correlations between those channels not only enhances the representation

capacity of each single channel, but also prevents the information loss, resulting in distributed alignment with minor changes to the semantic content.

As for segmentation model, the segmentation results of each category are obtained by multiplying all the channels by their corresponding weights and then adding them up. The weight value determines the influence of channel on the category. Hence, to identify each channel belongs to which category, we utilize the classifier from the SAN module. As shown in Fig. 3 (c), for each category, there are K weights corresponding to K channels, *i.e.* $\{w^1_{\text{class } c}, w^2_{\text{class } c}, ..., w^K_{\text{class } c}\}$ where $c \in \{1, ..., C\}$. After turning them into absolute values. We rank the weights of each category from the largest to the smallest. Then in each category, the first $\frac{K}{C}$ weight indexes are selected. For the sake of clarity, we use $\mathcal{I} \in \mathbb{R}^{C \times \frac{K}{C}}$ to denote the all selected indexes, where $\mathcal{I}(i, j)$ represents the j-th selected index of i-th category, $i \in \{1, ..., C\}$ and $j \in \{1, ..., \frac{K}{C}\}$. We arrange $\frac{K}{C}$ groups and allocate C different-category-related channels for each group. Specifically, each channel is weighted by its corresponding classifier weight before grouping, aiming to execute adaptive whitening transformation. Therefore, the m-th group of n-th sample: $\mathbf{G}^m_n$ in Eq. (14) can be modified as $\bar{\mathbf{G}}^m_n$:

$$\bar{\mathbf{G}}^m_n = [\widetilde{\mathbf{F}}_{n,\mathcal{I}(1,m)} \cdot w^{\mathcal{I}(1,m)}_{\text{class1}}; \widetilde{\mathbf{F}}_{n,\mathcal{I}(2,m)} \cdot w^{\mathcal{I}(2,m)}_{\text{class2}}; ...; \widetilde{\mathbf{F}}_{n,\mathcal{I}(C,m)} \cdot w^{\mathcal{I}(C,m)}_{\text{classC}}]. \tag{16}$$

Correspondingly, our whitening loss is formulated as:

$$\mathcal{L}_{\text{SAW}} = \frac{1}{N} \sum_{n=1}^{N} \sum_{m=1}^{\frac{K}{C}} ||\Psi(\bar{\mathbf{G}}^m_n) - \mathbf{I}||_1. \tag{17}$$

Note that operations of SAW module do not change the features of main network in forward pass. Therefore, different from SAN module, SAW is only applied during training.

5. Experiments

5.1. Datasets Description

Synthetic Datasets. GTA5 [53] is a synthetic image dataset, which is collected by using GTA-V game engine. It contains 24966 images with a resolution of 1914 × 1052 along with their pixel-wise semantic labels. SYNTHIA [54] is a large synthetic dataset with pixel-level semantic annotations. The subset SYNTHIA-RANDCITYSCAPES [54] is used in our experiments which contains 9400 images with a high resolution of 1280 × 760.

Real-World Datasets. Cityscapes [10] is a high resolution dataset (*i.e.* 2048 × 1024) of 5000 vehicle-captured urban street images taken from from 50 different cities primarily in Germany. BDDS [65] contains thousands of real-world dashcam video frames with accurate pixel-wise annotations, where 10000 images are provided with a resolution of 1280 × 720. Mapillary [45] contains 25000 images with diverse resolutions. The annotations contain 66 object categories, but only 19 categories overlap with others.

5.2. Implementation Details

We initialize the weights of the feature extractor module with an ImageNet [11] pre-trained model. We use SGD [26] optimizer with with an initial learning rate of 5e-4, a batch size of 2, a momentum of 0.9 and a weight decay of 5e-4. Besides, we follow the polynomial learning rate scheduling [6] with the power of 0.9. We train model for 200000 iterations. All datasets have 19 common categories, thus the parameter C defined in both SAN and SAW is set to 19. However, since SAW arranges $\frac{K}{C}$ groups, C can only be 2, 4, 8 or 16 to make K (channel number) divisible by C. Based on the results of ablation study on parameter C (Sec. 5.5), we set $C = 4$ in both SAN and SAW.

We implement our method on PyTorch [49] and use a single NVIDIA RTX 3090 GPU for our experiments. Following previous works, we use PASCAL VOC Intersection over Union (IoU) [13] as the evaluation metric.

5.3. Comparison with DG and DA methods

For brevity, we use G, S, C, B and M to denote GTA5 [53], SYNTHIA [54], Cityscapes [10], BDDS [65] and Mapillary [45], respectively. We extensively evaluate our method with different backbones including VGG-16 [57], ResNet-50 and ResNet-101 [22]. We repeat each experiment three times, and report the average results. Unlike existing synthetic-to-real generalization approaches, we propose to evaluate our model from an arbitrary domain to other unseen domains. Therefore, we conduct comprehensive experiments on five generalization settings, from **(1)** G to C, B, M & S; **(2)** S to C, B, M & G; **(3)** C to G, S, B & M; **(4)** B to G, S, C & M; **(5)** M to G, S, C & B. Our results reported in Tab. 1 on the first three settings, and Appendix A on the remaining 2 settings, suggest that our model consistently gains the best performance across all settings and backbones. Compared to the the second best results (see underlined values), our method shows a large improvement. Visual samples for qualitative comparison are given in Fig. 4. Remarkably, our method performs favorably in comparison to the methods that have access to the target domain data, see results in Appendix B.

5.4. Source Domain Performance Decay Analysis

A common pitfall of the domain generalization methods is that their performance degrades on the source domain. To compare different methods for this aspect, we evaluate them on the test set of the source domain in Tab. 2. The results suggest that our method largely retains performance on the source domain, and performs comparably with the model trained without domain-generalization (Baseline in Tab. 2).

Table 1. Performance comparison in terms of mIoU (%) between Domain Generalization methods. The best and second best results are **highlighted** and underlined, respectively. † denotes our re-implemention of the respective method. G, C, B, M and S denote GTA5 [53], Cityscapes [10], BDDS [65], Mapillary [45] and SYNTHIA [54], respectively.

Methods	Publication	Backbone	Train on GTA5 (G)				Train on SYNTHIA (S)				Train on Cityscapes (C)			
			→C	→B	→M	→S	→C	→B	→M	→G	→B	→M	→G	→S
Baseline			28.89	25.44	26.87	25.72	22.90	23.15	20.81	25.23	43.15	50.91	41.53	22.36
IBN† [47]	ECCV 2018		31.25	31.68	33.27	26.45	31.68	28.34	29.97	26.03	45.55	53.63	43.64	24.78
SW [48]	ICCV 2019		35.70	27.11	27.98	26.65	28.00	26.80	24.70	25.82	45.37	53.02	42.79	23.97
DRPC [66]	ICCV 2019	VGG-16	36.11	31.56	32.25	26.89	35.52	29.45	32.27	26.38	46.86	55.83	43.98	24.84
GTR† [51]	TIP 2021		36.10	32.14	34.32	26.45	36.07	31.57	30.63	26.93	45.93	54.08	43.72	24.13
ISW† [9]	CVPR 2021		34.36	33.68	34.62	26.99	36.21	31.94	31.88	26.81	47.26	54.21	42.08	24.92
Ours			**38.21**	**36.30**	**36.87**	**28.45**	**38.36**	**34.32**	**33.23**	**27.94**	**49.19**	**56.37**	**45.73**	**26.51**
Baseline			29.32	25.71	28.33	26.19	23.18	24.50	21.79	26.34	45.17	51.52	42.58	24.32
IBN [47]	ECCV 2018		33.85	32.30	37.75	27.90	32.04	30.57	32.16	26.90	48.56	57.04	45.06	26.14
SW [48]	ICCV 2019		29.91	27.48	29.71	27.61	28.16	27.12	26.31	26.51	48.49	55.82	44.87	26.10
DRPC [66]	ICCV 2019	ResNet-50	37.42	32.14	34.12	28.06	35.65	31.53	32.74	28.75	49.86	56.34	45.62	26.58
GTR† [51]	TIP 2021		37.53	33.75	34.52	28.17	36.84	32.02	32.89	28.02	50.75	57.16	45.79	26.47
ISW [9]	CVPR 2021		36.58	35.20	40.33	28.30	35.83	31.62	30.84	27.68	50.73	58.64	45.00	26.20
Ours			**39.75**	**37.34**	**41.86**	**30.79**	**38.92**	**35.24**	**34.52**	**29.16**	**52.95**	**59.81**	**47.28**	**28.32**
Baseline			30.64	27.82	28.65	28.15	23.85	25.01	21.84	27.06	46.23	53.23	42.96	25.49
IBN† [47]	ECCV 2018		37.42	38.28	38.28	28.69	34.18	36.63	36.19	28.15	50.22	58.42	46.33	27.57
SW† [48]	ICCV 2019		36.11	36.56	32.59	28.43	31.60	35.48	29.31	27.97	50.10	56.16	45.21	27.18
DRPC [66]	ICCV 2019	ResNet-101	42.53	38.72	38.05	29.67	37.58	34.34	34.12	29.24	51.49	58.62	46.87	28.96
GTR [51]	TIP 2021		43.70	39.60	39.10	29.32	39.70	35.30	36.40	28.71	51.67	58.37	46.76	29.07
ISW† [9]	CVPR 2021		42.87	38.53	39.05	29.58	37.21	33.98	35.86	28.98	50.98	59.70	46.28	28.43
Ours			**45.33**	**41.18**	**40.77**	**31.84**	**40.87**	**35.98**	**37.26**	**30.79**	**54.73**	**61.27**	**48.83**	**30.17**

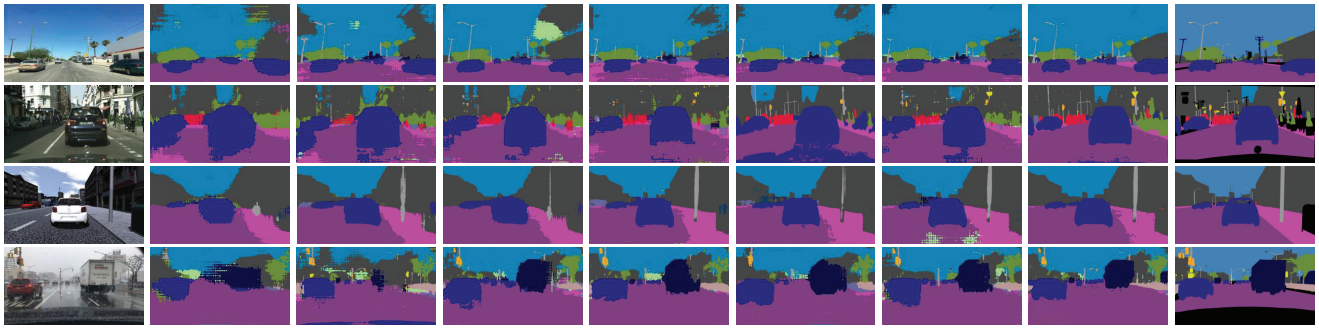


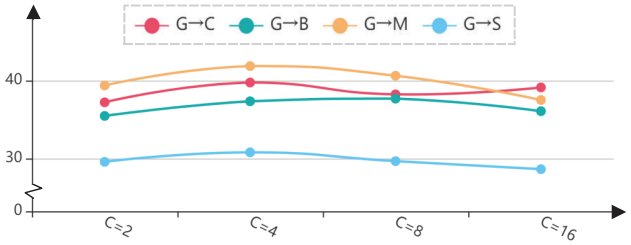
Figure 4. Visual comparison with different Domain Generalization methods on unseen domains *i.e*. Cityscapes [10], BDDS [65], Mapillary [45] and SYNTHIA [54], with the model trained on GTA5 [53]. The backbone network is ResNet-50.

5.5. Ablation Study

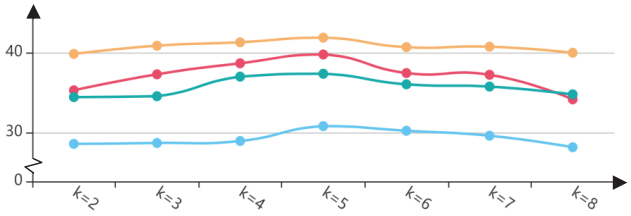
SAN and SAW. We investigate the individual contribution of SAN and SAW modules towards overall performance. Tab. 3 shows the mIoU improvement on ResNet-50 once we progressively integrate SAN and SAW. Experiments are conducted for generalization from GTA5 (G) to the other four datasets *i.e*. Cityscapes (C), BDDS (B), Mapillary (M) and SYNNTHIA(S). In our case, each of them helps boost the generalization performance by a large margin. Specifically, we observe that SAN and SAW greatly achieve an average improvement of 8.71% and 7.93%, respectively. We further observe that the models with only SAW show slightly weaker generalization capacity than those with only SAN. This is because without category-level centering of SAN, SAW performs distributed alignment, which may lead to incorrect feature matching between different categories. Therefore, our SAW module is mainly proposed to complement SAN in an integrated approach. As shown in Tab. 3, the best performance is achieved with a combination of both SAN and SAW.

CFR block in SAN. To demonstrate the effectiveness of the Category-level Feature Refinement (CFR) in SAN, we conducted an ablation experiment by removing the CFR block. As shown in Tab. 4 (top row), model without CFR consistently performs worse than model with SAN. With the help of CFR, SAN achieves an average gain of 2.30%, which demonstrates its usefulness.

Category-related Grouping in SAW. We perform ablations on different grouping strategies to verify their contributions. We conduct experiments on baseline with IW, GIW and SAW, respectively. The architectures of these



(a) mIoU Performance According to C



(b) mIoU Performance According to k

Figure 5. Change in performance with hyper-parameters: C and k. The experimental backbone is ResNet-50.

three models are illustrated in Fig. 3. As shown in Tab. 4 (bottom row), by adopting the general grouping operation, GIW shows significant improvement compared to model with IW. When applying our proposed SAW which performs category-related grouping, the performance of network improves to 37.54%, 34.97%, 39.85% & 28.46% on G $\rightarrow$ C, B, M & S. This confirms the effectiveness of category-related grouping in SAW.

Hyper-Parameter Analysis. C decides on the number of categories for semantic-aware feature alignment in both SAN and SAW. Since we rank the categories according to the their respective proportions in training data, only the first C categories are selected to handle category-level feature matching. To ensure that the channel dimension of feature maps is divisible by C, we can only set $C \in \{2, 4, 6, 8, 16\}$. Fig. 5 (a) shows compares results for different values, suggesting the optimal C is 4. In CFR block, we adopt k-means clustering to separate feature elements into category region and background. For searching the optimal parameter k, we only choose the first cluster whose cluster center is highest as the category region. As shown in Fig. 5 (b), model performs best when adopting k=5.

6. Conclusion and limitations

In this manuscript, we present a domain generalization approach to address the out-of-domain generalization for semantic segmentation. We propose two novel modules: Semantic-Aware Normalization (SAN) and Semantic-Aware Whitening (SAW), which sequentially perform category-level center alignment and distributed alignment to achieve both domain-invariant and discriminative features. Comprehensive experiments demonstrate the effectiveness of SAN and SAW with state-of-the-art performance in both domain generalization and domain adaptation. Although the method effectively eliminates feature differences caused by style variations on source domain, the extracted style-invariant features may still contain source-domain specific cues, leading to significant performance delta between the source and target domain. Some interesting future directions to close this gap include learning how to model domain shift (meta-learning), integrating multiple domain-specific neural networks (ensemble learning) and teaching a model to perceive generic features regardless of the target task (disentangled representation learning).

The datasets used in the paper lack diversity and have biases as they are mostly captured in the developed world. Beyond that, this paper have no ethical problem including personally identifiable information, human subject experimentation and military application.

Acknowledgement: This work was partially supported by the National Natural Science Foundation of China (No. 61972435, U20A20185). M. Hayat is supported by ARC DECRA fellowship DE200101100 .

Table 2. Comparison of different methods for performance drop on source domain. The performance drop (⇓) is obtained with respect to the baseline. The best and second best values are **highlighted** and underlined, respectively. The network backbone is ResNet-50.

Methods	GTA5	SYNTHIA	Cityscapes	BDDS	Mapillary
Baseline	73.95	70.84	77.93	79.67	70.49
IBN† [47]	⇓1.05	⇓2.40	⇓1.38	⇓1.30	⇓2.25
SW [48]	⇓0.45	⇓1.71	⇓0.63	⇓1.52	⇓0.91
DRPC [66]	⇓2.37	⇓3.63	⇓2.72	⇓2.46	⇓3.05
GTR† [51]	⇓2.07	⇓2.11	⇓3.25	⇓3.64	⇓4.17
ISW† [9]	⇓1.85	⇓1.28	⇓1.52	⇓1.83	⇓1.23
Ours	**⇓0.19**	**⇓0.08**	**⇓0.06**	**⇓0.23**	**⇓0.34**

Table 3. Ablation analysis of SAN (Sec. 4.1) and SAW (Sec. 4.2).

Methods	SAN	SAW	Train on GTA5 (G)			
			C	B	M	S
Baseline			29.32	25.71	28.33	26.19
+ SAN	✓		38.92	36.43	40.11	28.91
+ SAW		✓	37.54	34.97	39.85	28.46
All	✓	✓	**39.75**	**37.34**	**41.86**	**30.79**

Table 4. Ablation of different blocks in SAN and SAW

Methods	Train on GTA5 (G)			
	C	B	M	S
Baseline	29.32	25.71	28.33	26.19
Baseline + SAN (w/o CFR)	35.51	33.24	38.68	27.76
Baseline + SAN	**38.92**	**36.43**	**40.11**	**28.91**
Baseline + IW	33.19	31.27	30.55	26.55
Baseline + GIW	35.76	32.88	36.95	27.24
Baseline + SAW	**37.54**	**34.97**	**39.85**	**28.46**

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis & Machine Intelligence (TPAMI)*, 39(12):2481–2495, 2017. 2

[2] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 31:998–1008, 2018. 3

[3] Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. Analysis of representations for domain adaptation. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 19:137, 2007. 2

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 3431–3440, 2015. 2

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis & Machine Intelligence (TPAMI)*, 40(4):834–848, 2017. 2

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 2, 6

[7] Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7892–7901, 2018. 2

[8] 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 (CVPR)*, pages 10639–10647, 2019. 5

[9] 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 (CVPR)*, pages 11580–11590, 2021. 2, 3, 7, 8

[10] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016. 6, 7

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 6

[12] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 32:6450–6461, 2019. 3

[13] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision (IJCV)*, 111(1):98–136, 2015. 6

[14] Chuang Gan, Tianbao Yang, and Boqing Gong. Learning attributes equals multi-source domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 87–97, 2016. 3

[15] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1180–1189, 2015. 2

[16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *the Journal of Machine Learning Research (JMLR)*, 17(1):2096–2030, 2016. 2

[17] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012. 2

[18] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2551–2559, 2015. 3

[19] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. Dlow: Domain flow for adaptation and generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2477–2486, 2019. 3

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017. 2

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 6

[23] 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 *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1989–1998, 2018. 2

[24] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016. 2

[25] Wonsuk Kim and Junhee Seok. Indoor semantic segmentation for robot navigating on mobile. In *Proceedings of the IEEE International Conference on Ubiquitous and Future Networks (ICUFN)*, pages 22–25, 2018. 2

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 25:1097–1105, 2012. 6

[27] 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 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 61–71, 2021. 2

[28] Yinjie Lei, Duo Peng, Pingping Zhang, Qiuhong Ke, and Haifeng Li. Hierarchical paired channel fusion network for street scene change detection. *IEEE Transactions on Image Processing*, 30:55–67, 2020. 2

[29] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 3

[30] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1446–1455, 2019. 3

[31] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5400–5409, 2018. 3

[32] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *arXiv preprint arXiv:1705.08086*, 2017. 2, 3, 5

[33] 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. 3

[34] Yiying Li, Yongxin Yang, Wei Zhou, and Timothy Hospedales. Feature-critic networks for heterogeneous domain generalization. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 3915–3924, 2019. 3

[35] 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 (CVPR)*, pages 6936–6945, 2019. 3

[36] Hao Liu, Yulan Guo, Yanni Ma, Yinjie Lei, and Gongjian Wen. Semantic context encoding for accurate 3d point cloud segmentation. *IEEE Transactions on Multimedia*, 23:2045–2055, 2020. 2

[37] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015. 2

[38] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 97–105, 2015. 2

[39] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. *arXiv preprint arXiv:1602.04433*, 2016. 2

[40] Yanni Ma, Yulan Guo, Hao Liu, Yinjie Lei, and Gongjian Wen. Global context reasoning for semantic segmentation of 3d point clouds. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2931–2940, 2020. 2

[41] Ryusuke Miyamoto, Yuta Nakamura, Miho Adachi, Takeshi Nakajima, Hiroki Ishida, Kazuya Kojima, Risako Aoki, Takuro Oki, and Shingo Kobayashi. Vision-based road-following using results of semantic segmentation for autonomous navigation. In *Proceedings of the IEEE International Conference on Consumer Electronics (ICCE)*, pages 174–179, 2019. 2

[42] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5715–5725, 2017. 3

[43] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 10–18, 2013. 3

[44] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4500–4509, 2018. 2

[45] 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/CVF International Conference on Computer Vision (ICCV)*, pages 4990–4999, 2017. 6, 7

[46] 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 (CVPR)*, pages 3764–3773, 2020. 2

[47] 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, 7, 8

[48] 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 (ICCV)*, pages 1863–1871, 2019. 2, 3, 7, 8

[49] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban

Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *Proceedings of the Advances in Neural Information Processing Systems Workshops (NuerIPS Workshops)*, 2017. 6

[50] Duo Peng, Yinjie Lei, Wen Li, Pingping Zhang, and Yulan Guo. Sparse-to-dense feature matching: Intra and inter domain cross-modal learning in domain adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7108–7117, 2021. 2

[51] 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 (TIP)*, 30:6594–6608, 2021. 2, 3, 7, 8

[52] Mohammad Mahfujur Rahman, Clinton Fookes, Mahsa Baktashmotlagh, and Sridha Sridharan. Correlation-aware adversarial domain adaptation and generalization. *Pattern Recognition (PR)*, 100:107124, 2020. 3

[53] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 102–118, 2016. 6, 7

[54] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3234–3243, 2016. 6, 7

[55] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3723–3732, 2018. 2

[56] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3752–3761, 2018. 2

[57] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[58] Kihyuk Sohn, Sifei Liu, Guangyu Zhong, Xiang Yu, Ming-Hsuan Yang, and Manmohan Chandraker. Unsupervised domain adaptation for face recognition in unlabeled videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3210–3218, 2017. 2

[59] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4068–4076, 2015. 2

[60] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7167–7176, 2017. 2

[61] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6924–6932, 2017. 2, 3

[62] 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 (CVPR)*, pages 2517–2526, 2019. 2

[63] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3684–3692, 2018. 2

[64] Xiaoqing Ye, Jiamao Li, Hexiao Huang, Liang Du, and Xiaolin Zhang. 3d recurrent neural networks with context fusion for point cloud semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 403–417, 2018. 2

[65] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2(5):6, 2018. 6, 7

[66] 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 (ICCV)*, pages 2100–2110, 2019. 2, 3, 7, 8

[67] Pingping Zhang, Wei Liu, Yinjie Lei, and Huchuan Lu. Semantic scene labeling via deep nested level set. *IEEE Transactions on Intelligent Transportation Systems*, 22(11):6853–6865, 2020. 2

[68] Pingping Zhang, Wei Liu, Yinjie Lei, Hongyu Wang, and Huchuan Lu. Deep multiphase level set for scene parsing. *IEEE Transactions on Image Processing*, 29:4556–4567, 2020. 2

[69] Pingping Zhang, Wei Liu, Yinjie Lei, Hongyu Wang, and Huchuan Lu. Rapnet: Residual atrous pyramid network for importance-aware street scene parsing. *IEEE Transactions on Image Processing*, 29:5010–5021, 2020. 2

[70] Pingping Zhang, Wei Liu, Hongyu Wang, Yinjie Lei, and Huchuan Lu. Deep gated attention networks for large-scale street-level scene segmentation. *Pattern Recognition*, 88:702–714, 2019. 2

[71] Rui Zhang, Sheng Tang, Yongdong Zhang, Jintao Li, and Shuicheng Yan. Scale-adaptive convolutions for scene parsing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2031–2039, 2017. 2

[72] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2020–2030, 2017. 2

[73] Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Fully convolutional adaptation networks for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6810–6818, 2018. 2

[74] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 561–578, 2020. 3

[75] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. 2, 3