DIRL: Domain-Invariant Representation Learning for Generalizable Semantic Segmentation

Qi Xu[1], Liang Yao[2], Zhengkai Jiang[2], Guannan Jiang[3], Wenqing Chu[2], Wenhui Han[4], Wei Zhang[3], Chengjie Wang[2], Ying Tai[2]*

[1]Shanghai Jiao Tong University, Shanghai, China
[2]Tencent Youtu Lab, Shanghai, China
[3]Contemporary Amperex Technology Co., Limited, Shanghai, China
[4]Fudan University, Shanghai, China
txxqsh@sjtu.edu.cn, {liangyao, zhengkjiang, jasoncjwang, yingtai}@tencent.com, wqchu16@gmail.com, 19210980110@fudan.edu.cn, {jianggn, zhangwei}@catl.com

Abstract

Model generalization to the unseen scenes is crucial to real-world applications, such as autonomous driving, which requires robust vision systems. To enhance the model generalization, domain generalization through learning the domain-invariant representation has been widely studied. However, most existing works learn the shared feature space within multi-source domains but ignore the characteristic of the feature itself (e.g., the feature sensitivity to the domain-specific style). Therefore, we propose the Domain-invariant Representation Learning (DIRL) for domain generalization which utilizes the feature sensitivity as the feature prior to guide the enhancement of the model generalization capability. The guidance reflects in two folds: 1) Feature re-calibration that introduces the Prior Guided Attention Module (PGAM) to emphasize the insensitive features and suppress the sensitive features. 2): Feature whiting that proposes the Guided Feature Whiting (GFW) to remove the feature correlations which are sensitive to the domain-specific style. We construct the domain-invariant representation which suppresses the effect of the domain-specific style on the quality and correlation of the features. As a result, our method is simple yet effective, and can enhance the robustness of various backbone networks with little computational cost. Extensive experiments over multiple domains generalizable segmentation tasks show the superiority of our approach to other methods.

Introduction

Recently deep learning-based methods (Chen et al. 2017; Lin et al. 2017; Zheng et al. 2021) have obtained great progress in semantic segmentation, which greatly benefits from large-scale densely-annotated training data. However, when applying these models trained on the labeled dataset (source domain) to the unlabeled dataset (target domain), the performance drops significantly due to the huge domain gap. Therefore, how to reduce the domain gap to improve the model performance in the target domain has become a longstanding challenge in computer vision.

To tackle this challenge, Domain Adaptation (DA) (Tsai et al. 2018; Saito et al. 2018; Zou et al. 2018; Chu et al. 2019; Lee et al. 2020; Yu et al. 2021; Zhang et al. 2021) reduces the domain gap by aligning the data distribution between the source and target domains. However, DA requires to access the target domain which limits its application. In particular, this requirement is hard to be satisfied in the model adaptation to the real world since it is quite difficult to create a dataset that covers all real unseen scenes. Therefore, Domain Generalization (DG) has been widely studied to overcome this limitation. DG aims to improve the model generalization to the target domain without the target data in training. The essence of domain generalization is to learn domain-agnostic features. (Li et al. 2018b,a; Dou et al. 2019; Seo et al. 2020) learn the shared feature space within multi-source domains to construct domain-invariant representation. However, the question is *the characteristics of the feature itself (e.g., the feature sensitivity to the domain-specific style) are usually overlooked*. The feature sensitivity reflecting how likely the feature is domain-invariant can act as the useful prior knowledge to guide the learning of the domain-invariant representation. As shown in Fig. 1, when sending the images with the same content but different styles to

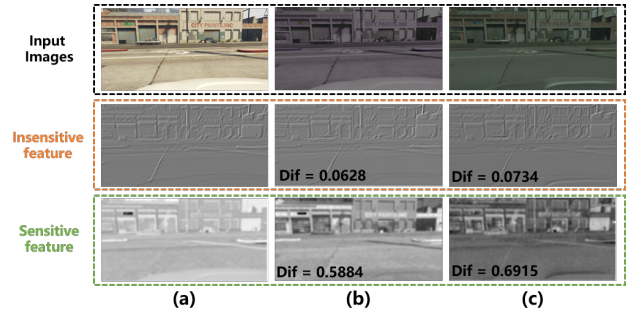


Figure 1: Illustration on feature sensitivity to the domain-specific style. We visualize some feature maps generated by the first residual group in ResNet50. The images share the same content but differ in style. (a) Original image. (b) Augmented image through photo-metric transform. (c) Augmented image through BDL-GAN (Li, Yuan, and Vasconcelos 2019). 'Dif' means the mean squared loss between the extracted features from the original and augmented images.

*Corresponding author.

the same network, some features are insensitive to the style while some are sensitive, which indicates that different features have different sensitivities to the domain-specific style.

In this paper, we explore the feature sensitivity to domain-specific style as the feature prior and propose a novel Domain-invariant Representation Learning (DIRL) for domain generalization in semantic segmentation. First, a Sensitivity-aware Prior Module (SAPM) is proposed to quantify the feature sensitivity as a guiding vector, which distinguishes the degree of feature change caused by the variance of style. Next, to embed the guidance of the feature prior into the network, we develop a Prior Guided Attention Module (PGAM) to re-calibrate the features under the guidance. The Sensitivity Guidance loss supervises the learning of the channel-wise attention weights to suppress the sensitive features and emphasize the insensitive features. In addition, we further adopt the feature whiting to promote model generalization, which has been proven effective in (Pan et al. 2019; Roy et al. 2019). However, directly adopting the feature whiting may eliminate the domain-specific style and domain-invariant content encoded in the features covariance in the meanwhile (Choi et al. 2021). Therefore, it is necessary to first decouple the features covariance into the domain-specific style and domain-invariant content, then selectively remove the domain-specific ones. Fortunately, the feature sensitivity to the domain-specific style is highly related to the features covariance sensitivity to the domain-specific style. The Guided Feature Whiting (GFW) is proposed to utilize the guidance of feature prior to decouple the features covariance, then the domain-specific ones are selectively removed. In general, our contributions are summarized as follows:

- To the best of our knowledge, this is the first work to explore feature sensitivity to the domain-specific style. We utilize the guidance of feature sensitivity to perform the feature re-calibration and feature whiting, which enhance the generalization capability.
- We propose a simple yet effective Domain-invariant Representation Learning (DIRL) algorithm, which consists of SAPM, PGAM, and GFW to realize the quantification and utilization of the feature prior (e.g., the feature sensitivity to the domain-specific style). These modules can be easily applied to existing models and significantly improve the generalization ability.
- We employ our method on multiple domains generalization tasks tailed to urban-scene segmentation. Extensive experiments show the superiority of DIRL over other existing approaches qualitatively and quantitatively.

Related Works

Domain Generalization

Domain Generalization (DG) aims to obtain a generalized model from the "known" source domain, which can perform well in various "unseen" target domains. Most DG methods can be broadly divided into two categories: Multi-source DG (Muandet, Balduzzi, and Schölkopf 2013; Ghifary et al. 2015; Li et al. 2018a; Seo et al. 2020; Bau et al. 2017; Mancini et al. 2018; Li et al. 2019) and single-source DG (Tobin et al. 2017; Yue et al. 2019; Qiao, Zhao, and Peng 2020; Choi et al. 2021; Huang et al. 2021).

Multi-source DG methods mainly learn a shared representation across multiple-source domains based on meta-learning (Li et al. 2018a), adversarial learning (Li et al. 2018a), metric learning (Dou et al. 2019) or auto-encoder (Seo et al. 2020). However, multiple domains are sometimes unavailable for training, and collecting multi domains is costly and labor-intensive. Hence, it's necessary to develop an effective learning paradigm for single-source DG.

Single-source DG methods can be divided into two categories: 1) Image-level based methods: Enrich the variation of synthetic images in the source domain through domain randomization (Tobin et al. 2017; Volpi et al. 2018; Yue et al. 2019; Huang et al. 2021). 2) Feature-level based methods: Introduce the instance normalization layers (Ulyanov, Vedaldi, and Lempitsky 2016) or feature whitening transformation (Pan et al. 2018; Seo et al. 2020; Choi et al. 2021) to eliminate the domain-specific style information. Different from the previous feature-level based single-source DG methods, we introduce the feature sensitivity to the domain-specific style as the feature prior to handle the domain generalization task, which is ignored in previous works but does improve the robustness of the learned representation.

Domain Adaptation for Semantic Segmentation

Domain adaptation methods aim to transfer the knowledge learned from the source domain to a specific target domain. Most DA methods can be divided into three categories: 1) Feature alignment through adversarial training (Valada et al. 2017; Vu et al. 2019). 2) Domain-specific knowledge learning through self-training (Zou et al. 2018, 2019). 3) Translating the source image to the target style to reduce the domain gap (Yang and Soatto 2020). DG is closely related to DA, but DG requires no access to the target domain. Moreover, a DG method can provide a good model initialization for DA.

Model Interpretability

It has been observed that many single hidden units can be aligned with human-explainable semantic concepts which are not explicitly taught to the network: Units have been found to detect objects, parts, textures, colors, scenes (Bau et al. 2017; Olah et al. 2018; Bau et al. 2018, 2020). Naturally, features extracted by the units matched well with the domain variant semantics, such as textures and colors, are sensitive to style, while features extracted by the units matched well with domain-invariant semantics, such as objects and parts, are insensitive to style. This inspires us to find out that the features have different sensitivities to the domain-specific style. Then we utilize the feature sensitivity to promote the domain generalization.

Method

The key of DIRL is the introduction of feature sensitivity to enhance the robustness of extracted features. As shown in Fig. 2, we first obtain the feature sensitivity through SAPM,

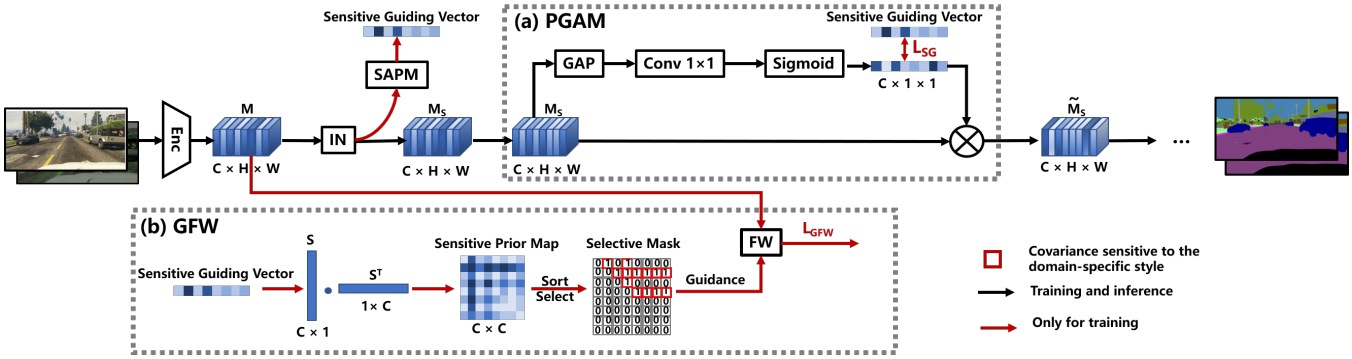


Figure 2: Overview of our proposed DIRL. (a) Prior Guided Attention Module. (b) Guided Feature Whiting. Enc: The encoder. FW: The feature whiting transform. GAP: The global average pooling. Conv: The convolution operation. 1×1: The kernel size. C×H×W: The tensor shape (depth, height, width). M: The intermediate feature map. M_s: The standardized feature map. $\tilde{M}_s$: The output feature map. L_{SG}: The Sensitivity Guidance loss. L_{GFW}: The Guided Feature Whiting loss.

then utilize the feature sensitivity to guide the feature recalibration and feature whiting, and finally feed the augmented features to the subsequent network to get the prediction result. Next, we will explain in detail each module and elaborate on our complete network structure.

Sensitivity-aware Prior Module (SAPM)

To quantify the feature sensitivity to the domain-specific style, we propose the Sensitivity-aware Prior Module as shown in Fig. 3. We think the domain-specific style information mainly reflects in color and blurriness, therefore we first simulate the style shift through photo-metric augmentation such as color jittering and Gaussian blurring.

Then we extract the corresponding feature maps by inferring from two input images, namely an original and a photo-metric transformed image, and calculate the differences between two different feature maps, which is defined as the difference vector d. Finally, we normalize each element of the difference vector into the same scale, i.e., between zero to one, to get the feature sensitivity, which is defined as the sensitive guiding vector s. Here it is worth noting that we can conveniently realize the calculation on the two inputs through the concatenating and splitting of the batch dimension. For brevity, we do not show the operation of the batch dimension in Fig. 2.

Formally, the feature difference vector $d \in R^{C\times1\times1}$ is:

$$d = GAP(L_2(M_b - M_b^{'})),$$

where $M_b, M_b^{'} \in \mathrm{R}^{1\times C\times H\times W}$ mean the feature maps for the original image and photo-metric transformed image. The normalization of the feature difference vector to get the sensitive guiding vector s is defined as:

$$s = \frac{d - Min(d)}{Max(d) - Min(d)} \in [0, 1],$$

where $Max(\cdot)$ and $Min(\cdot)$ compute the maximum and minimum value along the channel dimension. We denote each scalar in s as the Guiding Factor, where s_j is the Guiding Factor for channel j. The larger s_j is, the more sensitive the feature of channel j is to the style. It is worth mentioning that no additional trainable parameters or supervision are introduced in this module. We can apply SAPM at any position of the network to get the corresponding feature sensitivity.

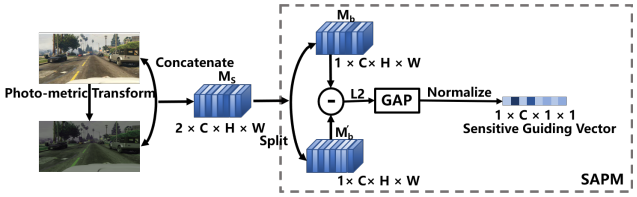


Figure 3: Illustration of Sensitivity-aware Prior Module. L2 indicates Euclidean norm.

Prior Guided Attention Module (PGAM)

After getting the feature prior, we need to utilize the guidance of the feature sensitivity to perform feature recalibration. We hope the network to learn a collection of per-channel attention weights, which realize the emphasis of insensitive features and the suppression of sensitive features. Therefore we introduce the PGAM incorporating a Sensitivity Guidance loss to learn the respective channel-wise attention weights under the guidance of feature sensitivity.

Specifically, we firstly adopt the global average pooling to aggregate feature maps across their spatial dimensions, then use the simple 1×1 convolution operation and a sigmoid activation to produce the channel-wise attention weights. These weights are applied to the original feature maps to generate the output of the PGAM, which can be fed directly into the subsequent layers of the network. To constrain the attention weights to reflect the feature sensitivity, we additionally introduce the Sensitivity Guidance loss L_{SG}.

Mathematically, we denote the attention weights as $w \in \mathrm{R}^{C\times1\times1}$. w has the same dimension with the sensitivity guidance vector s. We want the attention weights w and the sensitivity guidance vector s to be negatively correlated.

Therefore the Sensitivity Guidance loss L_{SG} is defined as:

$$L_{SG} = ||log(w)log(s) - 1||_2,$$

where $w, s \in [0, 1]$. The loss can constrain the attention weight to be close to 0 when the feature sensitivity is close to 1 for each channel.

Discussion. Why not directly adopt the feature sensitivity to re-calibrate the features? The reasons are two folds: 1) The obtainment of the feature sensitivity needs two inputs: original images and augmented images, which is not efficient in the inference stage. 2) The learnable attention weights are more flexible than the unlearnable feature sensitivity. The flexibility can promote the model generalization to the unseen scenes.

Guided Feature Whiting (GFW)

In addition to obtaining the re-calibrated features, we hope to further remove the effect of domain-specific style on the feature correlation through the whiting transform adopted in the pioneering work ISW (Choi et al. 2021). While performing the whitening transform, we firstly decouple the features covariance into the domain-specific and domain-invariant parts, then selectively suppress domain-specific ones.

The difference between our method and ISW (Choi et al. 2021) is that we utilize the feature sensitivity to guide decoupling while ISW utilizes the high-level statistics of the features covariance to perform decoupling. We argue the domain-specific features covariance mainly reflects in features covariance between the sensitive features and other features. Therefore, we adopt the sensitive guiding vector to generate Sensitive Prior Map (SPM) for the features covariance, where the larger the value of the position is, the more sensitive the position is to the style. Then we sort the values of every position and select the most sensitive positions to suppress them. Specifically, GFW contains three steps:

1) *Generate a covariance matrix Σ_s from a standardized feature map*. We send the feature map M to an instance normalization and get the normalized features M_s. Then the covariance matrix of the normalized features is calculated by

$$\Sigma_s = \frac{1}{HW}(M_s)(M_s)^T \in R^{C \times C}.$$

2) *Derive the selective mask for the covariance matrix from the sensitive guiding vector*. We use the sensitive guiding vector to generate the SPM which is defined as:

$$SPM = (S)(S)^T \in R^{C \times C}.$$

Then we select the top largest positions in the SPM to generate the Selective Mask (SM), which is a binary classifier to distinguish which position is sensitive to the domain-specific style. Since the covariance matrix is symmetric, SM only contains the strictly upper triangular part. The selective ratio α is set as 0.3 empirically.

3) *Adopt the SM to guide the feature whiting, which controls the selected covariance values to* 0. We use the Guided Feature Whiting loss L_{GFW} to remove the feature correlation in the selected positions, which is defined as

$$L_{GFW} = E[||\Sigma_s \bigodot SM||],$$

where E means the the arithmetic mean and SM indicates the generated selective mask.

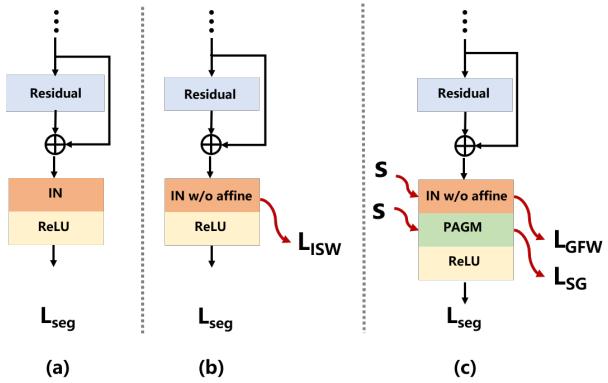


Figure 4: Architecture comparison with other methods: (a) IBN-Net mdule (Pan et al. 2018). (b) ISW module (Choi et al. 2021). (c) Our proposed layer which realizes the feature whiting and feature re-calibration under the guidance of the feature sensitivity s at the same time. L_{ISW} is Instance Selective Whitening loss proposed in (Choi et al. 2021).

Network Architecture in DIRL

Our design is inspired from IBN-Net (Pan et al. 2018) and ISW (Choi et al. 2021), as shown in Fig. 4. IBN-Net adopts the instance normalization to prevent over-fitting in the source domain and ISW further introduces whiting transform to solve DG. However, they both ignore the characteristics of the feature itself. We adopt a similar network architecture but introduce feature sensitivity as the guidance of feature re-calibration and feature whiting. The feature sensitivity serves as the useful prior knowledge for constructing the domain-invariant representation to enhance the model generalization. Specifically, we further add PGAM after the instance normalization and apply our proposed L_{GFW} and L_{SG} to the instance normalization layer and the PGAM, respectively. Our entire loss is described as:

$$L_{total} = L_{seg} + \lambda_1(\frac{1}{N}\sum_{i=1}^{N} L_{SG}^i) + \lambda_2(\frac{1}{N}\sum_{i=1}^{N} L_{GFW}^i),$$

where λ_1 and λ_2 are two constants balancing each loss, i denotes the layer index, and N is the number of applying this layer, L_{seg} means the segmentation loss which is defined as the pixel-wise cross entropy loss:

$$L_{seg} = -\sum_{i=1}^{M}\sum_{j=1}^{H\times W}\sum_{c=1}^{C} y_{ijc} log(p_{ijc}),$$

where M is the number of training images, H and W mean the image size, j denotes the pixel index, C represents the number of categories, c is the category index, $y_{ijc} \in \{0, 1\}$ is the one-hot vector representation of the ground-truth label and p_{ijc} is the predicted category probability.

Method	L_{SG}	L_{ISW}	L_{GFW}	mIoU		
				C	B	M
Baseline				28.95	25.14	28.18
DA				30.81	26.32	29.05
DU				33.57	28.74	30.24
Our Method	✓			36.60	30.66	33.55
	✓	✓		40.20	38.10	40.79
	✓		✓	**41.04**	**39.15**	**41.60**

Table 1: Abalation experiment for domain generalization task: GTAV→{Cityscapes, BDD and Mapillary} (using ResNet-50 as backbone) in mIoU. Notation: 'Baseline' means the DeepLabV3+ (Chen et al. 2017). 'DA' means the direct addition of PGAM to re-calibrate the features without the guidance of feature sensitivity. 'DU' means the direct utilization of the feature sensitivity to re-calibrate the features. L_{SG} means the Sensitivity Guidance loss. L_{ISW} means the Instance Selective Whitening loss (Choi et al. 2021). L_{GFW} denotes the Guided Feature Whiting loss.

Choice of α				
Value	0.2	0.3	0.4	0.5
Mean mIoU	38.30	**40.60**	39.07	38.11
Choice of λ_1				
Value	0.4	0.6	0.8	1.0
Mean mIoU	37.85	39.77	**40.60**	39.50
Choice of λ_2				
Value	0.2	0.4	0.6	0.8
Mean mIoU	38.79	38.94	**40.60**	39.57

Table 2: Performance with different parameters $\alpha, \lambda_1, \lambda_2$ in domain generalization task: GTAV→{Cityscapes, BDD and Mapillary}. Mean mIoU here is obtained in three datasts.

Experiments

In this section, we first introduce our used model and dataset. Then we explain our training details. After that, we illustrate the effectiveness of each component in our method through the ablation experiments. Finally, we present evaluation results to prove the effectiveness of our method on model generalization with comparisons to other methods.

Model and Dateset

Model. To illustrate the wide applicability of our proposed methods, We adopt DeepLabV3+ (Chen et al. 2017) with three backbones: ResNet (He et al. 2016), ShuffleNet (Ma et al. 2018) and MobileNet (Sandler et al. 2018) as the segmentation model, respectively.

Dataset. We evaluate the proposed algorithm on two challenging and important unsupervised domain generalization tasks: GTAV→{Cityscapes, BDD, Mapillary} and Cityscapes→{BDD, Synthia, GTAV} which involve two synthetic datasets and three real datasets.

Synthetic Dataset. GTAV (Richter et al. 2016) is a large-scale dataset containing $24,966$ high-resolution synthetic images. It contains $12,403$, $6,382$, and $6,181$ images of size 1914×1052 for training, validating, and testing respectively. It has 19 object categories compatible with Cityscapes. Synthia (Ros et al. 2016) consists of $9,400$ synthetic images with a resolution of 960×720, which shares 16 classes with the three target datasets.

Real Dataset. Cityscapes (Cordts et al. 2016) is a large semantic segmentation dataset, which is split into the training, validation, and testing parts with $2,975$, 500 and $1,525$ images respectively. BDD (Yu et al. 2020) is another real-world dataset that contains diverse urban driving scene images with the resolution of 1280×720. BDD provides $7,000$ images for training and $1,000$ images for validating. The last real-world dataset we adopt is Mapillary (Neuhold et al. 2017) consisting of $25,000$ high-resolution images with a minimum resolution of 1920×1080 collected from all around the world.

Training Details

We implement our method in Pytorch (Paszke et al. 2019). The optimizer is SGD with an initial learning rate of 0.01 and momentum of 0.9. Besides, we adopt the polynomial learning rate scheduling (Liu, Rabinovich, and Berg 2015) with the power of 0.9. We train all the models for 40K iterations with the batch size of 8. We adopt the color and positional augmentations such as color jittering, Gaussian blur, random cropping, random horizontal flipping, and random scaling with the range of $[0.5, 2.0]$ to avoid the model overfitting. For the photo-metric transformation in SAPM, we apply color jittering and Gaussian blurring. As shown in IBN-Net, earlier layers tend to encode the style information. Therefore we add the instance normalization layer and PGAM after the first three convolution groups.

Ablation Experiments

We examine each component of DIRL to find out how they contribute to the network generalization in semantic segmentation in Table 1. First, it is observed that the baseline model does not perform well due to the huge domain bias. Then, directly adopting the feature sensitivity or adding the PGAM to re-calibrate the features will bring in a certain performance improvement but is limited by the unlearnable feature sensitivity and the absence of feature guidance. The addition of the Sensitivity Guidance loss provides the learnable attention weights for feature re-calibration which brings in the significant improvement in the model generalization performance, especially in the task: GTAV→Cityscapes.

Next, we further compare two feature whiting losses in the last two columns. As we can see, Guided Feature Whiting loss obtains better performance because the feature sensitivity to the domain-specific style is closely related to the features covariance sensitivity to the domain-specific style, which demonstrates the importance of introducing the feature sensitivity as the guidance for domain generalization.

Sensitivity to Hyper-parameters. We further investigate the sensitivity of our method to the hyper-parameters $\alpha, \lambda_1, \lambda_2$ and show the results in Table 2. It can be seen

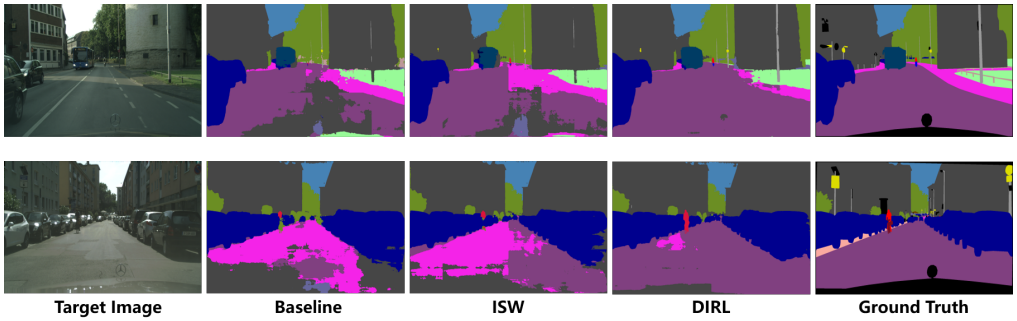


Figure 5: Qualitative illustration of domain generalizable semantic segmentation. DIRL introduces the feature sensitivity as the guidance to realize the domain-invariant representation learning which improves the segmentation performance.

Backbone	Method	C	B	M	Mean
ResNet50	Baseline	28.95	25.12	28.18	27.42
	SW	29.91	27.48	29.71	29.03
	IBN-Net	33.85	32.30	37.75	34.63
	IterNorm	31.81	32.70	33.88	32.80
	DPRC	37.42	32.14	34.12	34.56
	IW	33.21	32.67	37.35	34.41
	IRW	33.57	33.18	38.42	35.06
	ISW	36.58	35.20	40.33	37.37
	DIRL	**41.04**	**39.15**	**41.60**	**40.60**
ShuffleNet	Baseline	25.56	22.17	28.60	25.44
	IBN-Net	27.10	31.82	34.89	31.27
	ISW	30.98	32.06	35.31	32.78
	DIRL	**31.88**	**32.57**	**36.12**	**33.52**
MobileNet	Baseline	25.92	25.73	26.45	26.03
	IBN-Net	30.14	27.66	27.07	28.29
	ISW	30.86	30.05	30.67	30.53
	DIRL	**34.67**	**32.78**	**34.31**	**33.92**

Table 3: Domain generalization performance for the task: GTAV→{Cityscapes, BDD and Mapillary} in mIoU. Compared methods include SW (Pan et al. 2019), IBN-Net (Pan et al. 2018), IterNorm (Huang et al. 2019), DPRC (Yue et al. 2019), IW (Choi et al. 2021), IRW (Choi et al. 2021) and ISW (Choi et al. 2021).

Backbone	Method	B	S	G	Mean
ResNet50	Baseline	44.96	23.29	42.55	36.93
	SW	48.49	26.10	44.87	39.82
	IBN-Net	48.56	26.14	45.06	39.92
	IterNorm	49.23	25.98	45.73	40.31
	DPRC	N/A	N/A	N/A	N/A
	IW	48.19	25.81	45.21	39.74
	IRW	48.67	26.05	45.64	40.12
	ISW	50.73	26.20	45.00	40.64
	DIRL	**51.80**	**26.50**	**46.52**	**41.61**
ShuffleNet	Baseline	38.09	21.25	36.45	31.93
	IBN-Net	41.89	22.99	40.91	35.26
	ISW	41.94	22.82	40.17	34.98
	DIRL	**42.55**	**23.74**	**41.23**	**35.84**
MobileNet	Baseline	40.13	21.64	37.32	33.03
	IBN-Net	44.97	23.23	41.13	36.44
	ISW	45.17	22.91	41.17	36.42
	DIRL	**47.55**	**23.29**	**41.43**	**37.42**

Table 4: Domain generalization performance for the task: Cityscapes→{BDD, Synthia and GTAV} in mIoU. 'N/A' means the results are not reported in the paper.

Models	Params(M)	GFLOPS	Inference Time(ms)
Baseline	45.082	554.31	10.71
IBN-Net	45.083	554.31	10.18
ISW	45.081	554.31	10.50
DIRL	45.414	554.98	10.93

Table 5: Comparison of computational cost. Notation: the adopted backbone is ResNet50. The testing is performed with the image size of 2048×1024 on NVIDIA V100 GPU. The inference time is averaged over 500 times experiments to avoid the influence of volatility.

that the generalization performance firstly increases then decreases with the increase of three hyper-parameters, illustrating a bell shape curve. We empirically set the weight $\alpha, \lambda_1, \lambda_2$ as 0.3, 0.8, 0.6 to achieve the best performance.

Comparisons with State-of-the-Art Methods

Here we prove the superiorty of our method with other state-of-the-art DG methods. First, as shown in Table 3, our method outperforms other methods clearly and consistently across three different network backbones in the task from synthetic scenes to the real scenes. The superior segmentation performance is largely attributed to the introduction of the feature sensitivity, which guides the network to perform the feature re-calibration and feature whiting. DIRL provides robust representations which suppress the effect of the domain-specific style. Qualitative comparisons are provided in Fig. 5 to better illustrate the superiority of DIRL.

Then we further conduct the domain generalization task Cityscapes→{BDD, Synthia and GTAV} to provide the model generalization from the real scenes to the synthetic scenes and adverse scenes in Table 4. DIRL provides more reasonable predictions than other methods for the adverse conditions not included in Cityscapes, such as low illumina-

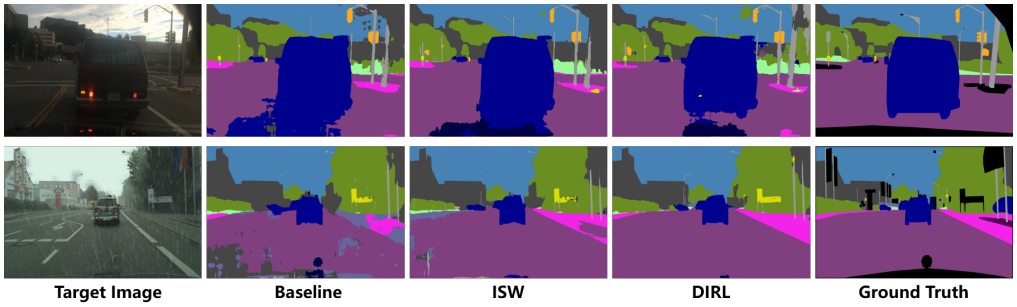


Figure 6: Segmentation results on adverse conditions in BDD (Yu et al. 2020) and RainyCityscapes (Hu et al. 2019) with the models trained on Cityscapes. Though Cityscapes does not contain scenes of adverse conditions, DIRL makes reasonable predictions in these cases.

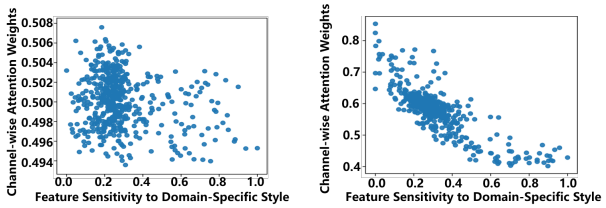
tion and rainy in Fig. 6. Though these types of scenes are not included in the training data, they are unavoidable and crucial for real-world applications (e.g., autonomous driving).

Computational Cost Analysis. Though our approach adds the additional module in the network, this brings in only a little additional computational cost. As shown in Fig. 4, our approach shares the similar architecture with other methods and only differs in the normalization layer and the introduction of PGAM. We report the number of parameters, GFLOPS, and inference time in Table 5, which proves the efficiency of DIRL.

Qualitative Analysis

Comparison of Channel-wise Attention Weights. To show how the feature sensitivity guides feature recalibration, we show the relation between the feature sensitivity to domain-specific style with the channel-wise attention weights before and after adding the sensitivity guiding loss in Fig. 7. It can be seen that the attention weights learned by the network itself have no obvious distinction and are almost around 0.5, while the guidance of feature sensitivity constrains the network to emphasize the insensitive features and suppress the sensitive features. Interestingly, the features with similar sensitivity have inconsistent attention weights. It seems after the network satisfies the guidance constraint, it further models the interdependencies among different channels, similar to (Hu, Shen, and Sun 2018).

Difference between ISW and GFW. ISW explores the sensitivity of *feature convariances* to style, while GFW further explores the sensitivity of *feature itself* to style. Sensitivity of the feature itself not only reflects the importance of the different feature parts in representation learning (for feature re-calibration), but also provides the guidance on decoupling sensitive and insensitive parts (for feature whiting), which both improve the robustness of learnt representation. GFW promotes more thorough decoupling between the sensitive and insensitive parts of the learnt feature than ISW. As shown in Fig. 8, the distribution of ISW is more averaged than GFW, which effectively proves the two parts has been separated well in GFW while not enough good in ISW. Better decoupling can avoid the affect of the domain-specific style on the learnt representation (Wu et al. 2021).



(a) without L_{SG} (b) with L_{SG}

Figure 7: The relation between the feature sensitivity with the channel-wise attention weights, which is induced by the PGAM in the third convolution group in ResNet50.

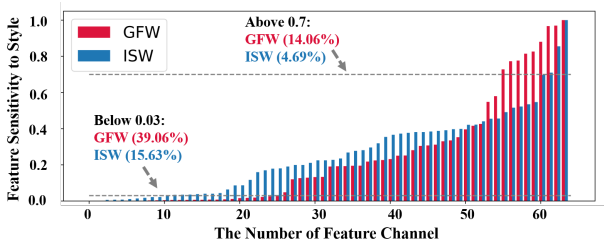


Figure 8: Comparison between GFW and ISW in feature sensitivity of the first convolution in ResNet50, which has been sorted to highlight differences.

Conclusion

This paper introduces the feature sensitivity to the domain-specific style as the prior knowledge to guide the feature recalibration and feature whiting, which promotes the learning of the domain-invariant representation. We show the potential of our proposed Domain-invariant Representation Learning (DIRL) in the urban segmentation, including the domain generalization from synthetic scenes to real scenes, and from general conditions to adverse conditions. In the future, we strive to improve the model generalization to promote the application of deep neural networks on outdoor scenes, such as autonomous driving.

References

Bau, A.; Belinkov, Y.; Sajjad, H.; Durrani, N.; Dalvi, F.; and Glass, J. 2018. Identifying and controlling important neurons in neural machine translation. *arXiv preprint arXiv:1811.01157*.

Bau, D.; Zhou, B.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6541–6549.

Bau, D.; Zhu, J.-Y.; Strobelt, H.; Lapedriza, A.; Zhou, B.; and Torralba, A. 2020. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117(48): 30071–30078.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(4): 834–848.

Choi, S.; Jung, S.; Yun, H.; Kim, J. T.; Kim, S.; and Choo, J. 2021. RobustNet: Improving Domain Generalization in Urban-Scene Segmentation via Instance Selective Whitening. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 11580–11590.

Chu, W.; Hung, W.-C.; Tsai, Y.-H.; Cai, D.; and Yang, M.-H. 2019. Weakly-supervised caricature face parsing through domain adaptation. In *The IEEE International Conference on Image Processing (ICIP)*, 3282–3286.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3213–3223.

Dou, Q.; Coelho de Castro, D.; Kamnitsas, K.; and Glocker, B. 2019. Domain generalization via model-agnostic learning of semantic features. *Advances in Neural Information Processing Systems (NeurIPS)*, 32: 6450–6461.

Ghifary, M.; Kleijn, W. B.; Zhang, M.; and Balduzzi, D. 2015. Domain generalization for object recognition with multi-task autoencoders. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2551–2559.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7132–7141.

Hu, X.; Fu, C.-W.; Zhu, L.; and Heng, P.-A. 2019. Depth-attentional features for single-image rain removal. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 8022–8031.

Huang, J.; Guan, D.; Xiao, A.; and Lu, S. 2021. Fsdr: Frequency space domain randomization for domain generalization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6891–6902.

Huang, L.; Zhou, Y.; Zhu, F.; Liu, L.; and Shao, L. 2019. Iterative normalization: Beyond standardization towards efficient whitening. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4874–4883.

Lee, S.; Hyun, J.; Seong, H.; and Kim, E. 2020. Unsupervised Domain Adaptation for Semantic Segmentation by Content Transfer. *arXiv preprint arXiv:2012.12545*.

Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. M. 2018a. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*.

Li, D.; Zhang, J.; Yang, Y.; Liu, C.; Song, Y.-Z.; and Hospedales, T. M. 2019. Episodic training for domain generalization. In *The Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1446–1455.

Li, H.; Pan, S. J.; Wang, S.; and Kot, A. C. 2018b. Domain generalization with adversarial feature learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5400–5409.

Li, Y.; Yuan, L.; and Vasconcelos, N. 2019. Bidirectional learning for domain adaptation of semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6936–6945.

Lin, G.; Milan, A.; Shen, C.; and Reid, I. 2017. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1925–1934.

Liu, W.; Rabinovich, A.; and Berg, A. C. 2015. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*.

Ma, N.; Zhang, X.; Zheng, H.-T.; and Sun, J. 2018. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, 116–131.

Mancini, M.; Bulo, S. R.; Caputo, B.; and Ricci, E. 2018. Best sources forward: domain generalization through source-specific nets. In *The IEEE international conference on image processing (ICIP)*, 1353–1357. IEEE.

Muandet, K.; Balduzzi, D.; and Schölkopf, B. 2013. Domain generalization via invariant feature representation. In *International Conference on Machine Learning (ICML)*, 10–18.

Neuhold, G.; Ollmann, T.; Rota Bulo, S.; and Kontschieder, P. 2017. The mapillary vistas dataset for semantic understanding of street scenes. In *The Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4990–4999.

Olah, C.; Satyanarayan, A.; Johnson, I.; Carter, S.; Schubert, L.; Ye, K.; and Mordvintsev, A. 2018. The building blocks of interpretability. *Distill*, 3(3): e10.

Pan, X.; Luo, P.; Shi, J.; and Tang, X. 2018. Two at once: Enhancing learning and generalization capacities via ibn-net. In *The Proceedings of the European Conference on Computer Vision (ECCV)*, 464–479.

Pan, X.; Zhan, X.; Shi, J.; Tang, X.; and Luo, P. 2019. Switchable whitening for deep representation learning. In *The Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1863–1871.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems (NeurIPS)*, 32: 8026–8037.

Qiao, F.; Zhao, L.; and Peng, X. 2020. Learning to learn single domain generalization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 12556–12565.

Richter, S. R.; Vineet, V.; Roth, S.; and Koltun, V. 2016. Playing for data: Ground truth from computer games. In *The Proceedings of the European Conference on Computer Vision (ECCV)*, 102–118. Springer.

Ros, G.; Sellart, L.; Materzynska, J.; Vazquez, D.; and Lopez, A. M. 2016. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3234–3243.

Roy, S.; Siarohin, A.; Sangineto, E.; Bulo, S. R.; Sebe, N.; and Ricci, E. 2019. Unsupervised domain adaptation using feature-whitening and consensus loss. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 9471–9480.

Saito, K.; Watanabe, K.; Ushiku, Y.; and Harada, T. 2018. Maximum classifier discrepancy for unsupervised domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3723–3732.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4510–4520.

Seo, S.; Suh, Y.; Kim, D.; Kim, G.; Han, J.; and Han, B. 2020. Learning to optimize domain specific normalization for domain generalization. In *The European Conference on Computer Vision (ECCV)*, 68–83. Springer.

Tobin, J.; Fong, R.; Ray, A.; Schneider, J.; Zaremba, W.; and Abbeel, P. 2017. Domain randomization for transferring deep neural networks from simulation to the real world. In *The international conference on intelligent robots and systems (IROS)*, 23–30. IEEE.

Tsai, Y.-H.; Hung, W.-C.; Schulter, S.; Sohn, K.; Yang, M.-H.; and Chandraker, M. 2018. Learning to adapt structured output space for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7472–7481.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2016. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*.

Valada, A.; Vertens, J.; Dhall, A.; and Burgard, W. 2017. AdapNet: Adaptive semantic segmentation in adverse environmental conditions. In *The IEEE International Conference on Robotics and Automation (ICRA)*.

Volpi, R.; Namkoong, H.; Sener, O.; Duchi, J.; Murino, V.; and Savarese, S. 2018. Generalizing to unseen domains via adversarial data augmentation. *arXiv preprint arXiv:1805.12018*.

Vu, T.-H.; Jain, H.; Bucher, M.; Cord, M.; and Pérez, P. 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2517–2526.

Wu, A.; Han, Y.; Zhu, L.; and Yang, Y. 2021. Instance-Invariant Domain Adaptive Object Detection via Progressive Disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Yang, Y.; and Soatto, S. 2020. Fda: Fourier domain adaptation for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4085–4095.

Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; and Darrell, T. 2020. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2636–2645.

Yu, F.; Zhang, M.; Dong, H.; Hu, S.; Dong, B.; and Zhang, L. 2021. DAST: Unsupervised Domain Adaptation in Semantic Segmentation Based on Discriminator Attention and Self-Training. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 10754–10762.

Yue, X.; Zhang, Y.; Zhao, S.; Sangiovanni-Vincentelli, A.; Keutzer, K.; and Gong, B. 2019. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *The Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2100–2110.

Zhang, P.; Zhang, B.; Zhang, T.; Chen, D.; Wang, Y.; and Wen, F. 2021. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 12414–12424.

Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P. H.; et al. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6881–6890.

Zou, Y.; Yu, Z.; Liu, X.; Kumar, B.; and Wang, J. 2019. Confidence regularized self-training. In *The IEEE International Conference on Computer Vision (ICCV)*, 5982–5991.

Zou, Y.; Yu, Z.; Vijaya Kumar, B.; and Wang, J. 2018. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *The European Conference on Computer Vision (ECCV)*, 289–305.