Style-Hallucinated Dual Consistency Learning: A Unified Framework for Visual Domain Generalization

Yuyang Zhao[1] · Zhun Zhong[2] · Na Zhao[3] · Nicu Sebe[4] · Gim Hee Lee[1]

Abstract

Domain shift widely exists in the visual world, while modern deep neural networks commonly suffer from severe performance degradation under domain shift due to poor generalization ability, which limits real-world applications. The domain shift mainly lies in the limited source environmental variations and the large distribution gap between source and unseen target data. To this end, we propose a unified framework, Style-HAllucinated Dual consistEncy learning (SHADE), to handle such domain shift in various visual tasks. Specifically, SHADE is constructed based on two consistency constraints, Style Consistency (SC) and Retrospection Consistency (RC). SC enriches the source situations and encourages the model to learn consistent representation across style-diversified samples. RC leverages general visual knowledge to prevent the model from overfitting to source data and thus largely keeps the representation consistent between the source and general visual models. Furthermore, we present a novel style hallucination module (SHM) to generate style-diversified samples that are essential to consistency learning. SHM selects basis styles from the source distribution, enabling the model to dynamically generate diverse and realistic samples during training. Extensive experiments demonstrate that our versatile SHADE can significantly enhance the generalization in various visual recognition tasks, including image classification, semantic segmentation, and object detection, with different models, i.e., ConvNets and Transformer.

Keywords Domain generalization · Style variation · Consistency learning · Visual recognition

Communicated by Oliver Zendel.

✉ Yuyang Zhao
yuyang.zhao@u.nus.edu

Zhun Zhong
zhunzhong007@gmail.com

Na Zhao
na_zhao@sutd.edu.sg

Nicu Sebe
sebe@disi.unitn.it

Gim Hee Lee
gimhee.lee@nus.edu.sg

[1] Department of Computer Science, National University of Singapore, Singapore, Singapore

[2] School of Computer Science, University of Nottingham, Nottingham, UK

[3] Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore, Singapore

[4] Department of Information Engineering and Computer Science, University of Trento, Trento, Italy

1 Introduction

With the development of deep neural networks and the introduction of abundant annotated data, fully-supervised methods have achieved remarkable success in various visual recognition tasks, including but not limited to image classification (He et al., 2016; Dosovitskiy et al., 2021; Liu et al., 2021), object detection (Ren et al., 2015; He et al., 2017; Carion et al., 2020) and semantic segmentation (Hoffman et al., 2016; Chen et al., 2018; Xie et al., 2021). These visual recognition tasks are the fundamental and crucial components of the computer vision world. However, such significant achievements heavily rely on the availability of large-scale annotated data, which are expensive and time-consuming to collect, especially for semantic segmentation and object detection. For example, it takes more than 1.5 h to annotate the semantic labels of a 1024×2048 driving scene (Cordts et al., 2016), and the time is even doubled for scenes under adverse weather (Sakaridis et al., 2021) and poor lighting conditions (Sakaridis et al., 2019). In addition, even though given abundant labeled training data, the
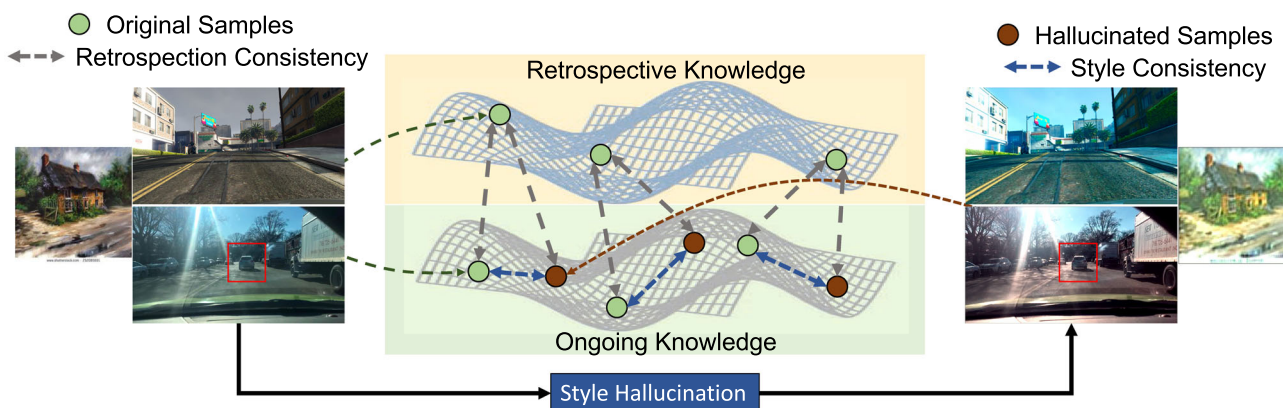
Fig. 1 Illustration of the proposed dual consistency constraints for three visual tasks. We generate hallucinated samples (brown circle) from the style hallucination module and then utilize the paired samples and general visual (retrospective) knowledge to learn style consistency (blue dash line) and retrospection consistency (gray dash line) (Color figure online)

significant performance of the deep learning model is limited to independent and identically distributed (*i.i.d*) datasets. Nevertheless, out-of-distribution (OOD) data that are totally unseen during training are inevitably in real-world applications, e.g., weather change, and the models commonly suffer from catastrophic performance degradation when facing unseen situations.

To alleviate the heavy annotation cost and distribution shift, domain generalization (DG) (Zhou et al., 2021a; Wu & Deng, 2022; Zhong et al., 2022) is introduced in the community. DG only leverages annotated source data to train a robust model that can cope with different unseen conditions. The source domain can be the annotated real-world data but can also be the synthetic data from pre-designed engine (Ros et al., 2016; Richter et al., 2016), where the latter can greatly reduce the annotation cost. In view of the practicality of DG, previous works have been independently investigating it in image classification (Zhou et al., 2021a), semantic segmentation (Zhong et al., 2022), and object detection (Wu & Deng, 2022). In this paper, we aim to propose a unified and versatile framework that is applicable to the above three visual recognition tasks.

The main challenge for DG is to cope with the significant domain shift between source and unseen target domains, which can be roughly divided into two aspects. First, the environmental variations and diversity in source data are very limited compared to those of unseen target data. Second, there exists a large distribution gap between the source and target data, *e.g.,* image styles and characteristics of objects. To learn the domain-invariant model that can address the domain shift, previous works mainly focus on three aspects: (1) designing tailor-made modules (Choi et al., 2021; Pan et al., 2018; Wu & Deng, 2022) to remove domain-specific information; (2) leveraging extra data to transfer source

data (Huang et al., 2021; Yue et al., 2019) to possible target styles for narrowing the distribution gap; (3) diversifying source data within the domain via style augmentation (Zhou et al., 2021a; Wang et al., 2021c) or adversarial perturbation (Zhong et al., 2022; Shankar et al., 2018). However, the removal of domain-specific information is not complete and explicit due to the lack of target information; the extra style transfer heavily relies on extra data, which are not always available in practice, and ignores the invariant representation within the source domain. Taking the above into account, in this paper, we follow the third paradigm to diversify samples in the source domain. In addition, we explicitly introduce two constraints to help the model effectively learn domain-invariant representation and narrow the domain gap.

To this end, we introduce a novel dual consistency learning framework that can jointly address the above two types of domain shift. As shown in Fig. 1, we introduce two consistency constraints, *style consistency* (SC) and *retrospection consistency* (RC). SC encourages the model to learn style invariant representation by forcing the consistency between the samples before and after style variation. RC aims at leading the model less overfitting to the source data with the help of general visual knowledge. Specifically, we leverage the ImageNet (Deng et al., 2009) pre-trained model which is available acquiescently in all DG models. The features from the pre-trained model can reflect the representation in the context of the general visual world and thus can serve as the guidance for the ongoing model to retrospect what the visual world looks like and to lead the model less overfitting to the source data.

Style diversifying is crucial for the success of dual consistency learning, and we adopt the style features, i.e., channel-wise mean and standard deviation, to generate new data. Compared with directly transferring the whole image

(e.g., CycleGAN (Zhu et al., 2017)), changing style features can maintain the pixel alignment to the utmost extent, which is better for pixel-level tasks like semantic segmentation. Previous works (Tang et al., 2021; Zhou et al., 2021a) commonly mix or swap styles within the source domain, which will generate more samples of the dominant styles (e.g., daytime in GTAV (Richter et al., 2016)). Nevertheless, it is not the best way since the target styles may be quite different from the dominant styles. To fully take advantage of all the source styles, we propose style hallucination module (SHM), which leverages C basis styles to represent the style space of C dimension and thus generate new styles. Ideally, the basis styles should be linearly independent so the linear combination of basis styles can represent all the source styles. However, many unrealistic styles that impair the model training are generated when we directly take C orthogonal unit vectors as the basis. To reconcile diversity and realism, we use farthest point sampling (FPS) (Qi et al., 2017) to select C styles from all the source styles as basis styles. Such basis styles contain many rare styles since rare styles are commonly far away from the dominant ones. With these basis styles that represent the style space in a better way, we utilize linear combination to generate new styles. To summarize, we propose the Style-HAllucinated Dual consistEncy learning (SHADE) framework for domain generalization in visual recognition tasks, and our contributions are as follows:

– We propose the dual consistency constraints for visual domain generalization, which can learn the style invariant representation among diversified styles and narrow the domain gap between source and target domains by retrospecting the general knowledge of the visual world.
– We propose the style hallucination module to generate new and diverse styles with representative basis styles, effectively facilitating dual consistency learning.
– The proposed SHADE is generic and can be applied to various domain generalized visual recognition tasks, including image classification, semantic segmentation, and object detection. In addition, SHADE can be applied to both ConvNets and Transformer to improve generalization ability.

An earlier and preliminary study of SHADE was published in ECCV 2022 (Zhao et al., 2022c). In comparison, this paper further introduces the following significant contributions. Specifically,

(1) we investigate the generalization ability of Transformer-based semantic segmentation model and demonstrate that many state-of-the-art DG methods cannot work well on this model but SHADE can still achieve significant performance;

(2) we propose the general formulation of dual consistency learning, which enables us to apply it to different visual tasks with a light modification;
(3) we demonstrate that SHADE is complementary to different state-of-the-art DG methods in image classification, leading to further improvement;
(4) we demonstrate that SHADE is applicable to domain generalized object detection to improve the robustness under the environmental change in the urban scenes;
(5) more ablation experiments are provided to further verify the effectiveness of each component of our method;
(6) visualizations of the three visual tasks are provided to better understand the effect of our SHM.

2 Related Work

Domain Generalization To tackle the performance degradation in the out-of-distribution conditions, domain generalization (DG) (Huang et al., 2021; Yue et al., 2019; Choi et al., 2021; Zhou et al., 2021a; Zhao et al., 2021; Gong et al., 2021; Yuan et al., 2022; Du et al., 2022) is introduced to learn a robust model with one or multiple source domains, aiming to perform well on unseen domains. The domain generalization methods can be divided into two parts based on the number of source domains, i.e., single-source DG and multi-source DG. In general, domain generalized model requires to learn the domain-invariant representation by smooth representation function with controlled complexity (Shui et al., 2022b). Thus, multi-source DG can explicitly learn the domain-invariant representations from multiple source domains (Li et al., 2018a, b; Zhao et al., 2021; Shui et al., 2021) or generating samples across source domains (Zhou et al., 2021a; Nuriel et al., 2021). Single-source DG is more complex, which can only investigate the robust representation from one source domain. The mainstream of single-source DG is to diversify the source sample with additional styles (Wang et al., 2021c; Tang et al., 2021) or perturbation (Qiao et al., 2020; Zhao et al., 2020).

In semantic segmentation, considering the expensive annotation cost, synthetic data are commonly adopted as the source domain in recent domain generalized semantic segmentation (DG-Seg) works. To address the domain shift, one mainstream of previous works (Huang et al., 2021; Yue et al., 2019) focuses on diversifying training data with real-world templates and learning the invariant representation from all the domains. Another mainstream aims at directly learning the explicit domain-invariant features within the source domain (Choi et al., 2021; Pan et al., 2018; Peng et al., 2022). IBN-Net (Pan et al., 2018) and ISW (Choi et al., 2021) leverage tailor-made instance normalization block and whitening transformation to reduce the impact of domain-specific information. More recently, style augmen-

tation based methods (Zhong et al., 2022; Lee et al., 2022) attract more attention in DG-Seg.

Domain generalized object detection (DG-Det) is a relatively new setting. The mainstream of DG-Det is disentangling domain-invariant features from the domain-specific features. Lin et al. (2021) leverage feature disentangle and representation reconstruction from two source domains to learn the generalizable detection model. Wu and Deng (2022) focus on the single-source DG-Det in the urban scenes, where they disentangle the domain-invariant and domain-specific information via cyclic operation and use self-distillation to improve the generalization ability.

In this paper, we focus on the visual domain generalization in image classification, semantic segmentation and object detection. We mainly focus on the single-source setting but also demonstrate the effectiveness of our method under the multi-source situations.

Consistency learning (CL) CL is adopted by many computer vision tasks and settings. One main stream is leveraging CL to exploit the unlabeled samples in semi-supervised learning (French et al., 2020; Tarvainen & Valpola, 2017), unsupervised domain adaptation (Zhou et al., 2022, 2021b; Chen et al., 2021b; Zheng & Yang, 2021, 2020, 2022; Chen et al., 2022) and novel class discovery (Zhao et al., 2022b; Fini et al., 2021; Roy et al., 2022; Zhong et al., 2021). CL is also applied to address the corruptions and perturbations (Hendrycks et al., 2020; Wang et al., 2021a) by maximizing the similarity between different augmentations. In addition, CL is also used in self-supervised learning (Chen et al., 2020; He et al., 2020) as the contrastive loss to utilize totally unlabeled data. We introduce CL into domain generalization, leading the model robust to various styles. We also leverage consistency with general visual knowledge to prevent the model from overfitting to the source data.

Style Variation Style features are widely explored in style transfer (Dumoulin et al., 2017; Huang & Belongie, 2017), which aims at changing the image style but maintaining the content. Inspired by this, recent domain generalization methods leverage the style features to generate diverse data of different styles to improve the generalization ability. Swapping (Tang et al., 2021; Zhao et al., 2022a) and mixing (Zhou et al., 2021a) existing styles within the source domains is an effective way and generating new styles (Wang et al., 2021c) by specially designed modules can also make sense. More recently, AdvStyle (Zhong et al., 2022) applies adversarial learning to the style features to generate diverse and hard samples. We also only leverage the styles within the source domain but take the relatively rare styles in the source domain into account, thus generating more diverse samples to improve generalization ability.

3 Methodology

Preliminary In domain generalization, one or multiple labeled source domains $\mathcal{S} = \{x_S^i, y_S^i\}_{i=i}^{N_S}$, where N_S is the number of source domains, are used to train a recognition model, which is deployed to unseen target domains $\mathcal{T}$ directly. In general, the source and target domains share the same label space $Y_S, Y_T \in (1, N_C)$ but belong to different distributions. The goal of this task is to improve the generalization ability of model in unseen domains using only the source data.

Overview To solve the above challenging problem, we propose the Style-HAllucinated Dual consistEncy learning (SHADE) framework, which is quipped with dual consistency constraints and a Style Hallucination Module (SHM). The dual consistency constraints effectively learn the domain-invariant representation and ameliorate the overfitting issue via general visual knowledge. SHM enriches the training samples by dynamically generating diverse styles, which catalyzes the advantage of dual consistency learning. The overall framework is shown in Fig. 2 and SHM is illustrated in Fig. 3.

3.1 Dual Consistency Constraints

In SHADE, we introduce two consistency learning constraints: (1) Style Consistency (SC) that aims at learning the consistent predictions across stylized samples. (2) Retrospection Consistency (RC) that focuses on narrowing the distribution shift between source data and general visual knowledge in the feature-level, which can alleviate the model overfitting to the source data.

Style Consistency (SC) Cross entropy constraint is one of the most important criterion for recognition tasks, focusing on the high-level semantic information for each class. Instead of the global representation for each class, logit pairing highlights the most invariant information across paired samples, which has demonstrated its effectiveness in learning adversarial samples (Hendrycks et al., 2020; Kannan et al., 2018). Inspired by this, we propose SC to ameliorate the style shift with logit pairing. To fulfill this, SC requires the style-diversified samples $\tilde{x}_S$ which share the same semantic content with the source samples x_S but of different styles. To generate stylized samples online and maintain the pixel-level semantic information, non-geometry style variation, e.g., MixStyle (Zhou et al., 2021a), CrossNorm (Tang et al., 2021) and the proposed style hallucination in Sect. 3.3, is used to obtain the style-diversified samples $\tilde{x}_S$. *Note that* we focus on three visual recognition tasks, including image classification, semantic segmentation and object detection. As shown in Fig. 2, the implementation of style consistency depends on the task. In general, we minimize the Jensen-
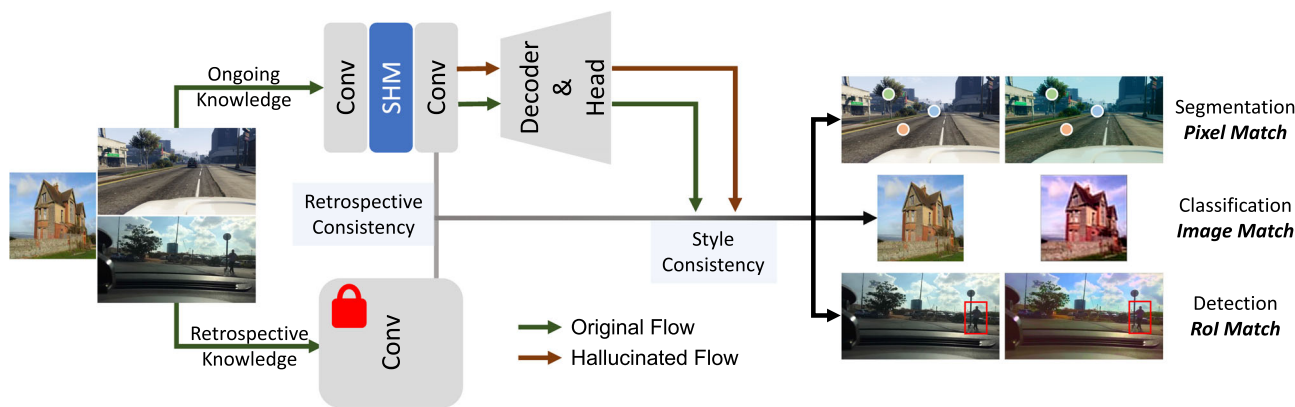
Fig. 2 Overview of the proposed style-hallucinated dual consistency learning (SHADE) framework. Training images are fed into the training model (ongoing knowledge) and the ImageNet pre-trained model (retrospective knowledge). The style hallucination module is inserted into a certain layer of the training model to generate stylized samples. Finally, the model is optimized by the dual consistency losses: style consistency $\mathcal{L}_{\text{SC}}$ and retrospection consistency $\mathcal{L}_{\text{RC}}$. Note that the standard task loss is also used and we omit it here for brevity

Shannon Divergence (JSD) between the posterior probability of the semantically aligned $\tilde{x}_S$ and x_S:

$$\mathcal{L}_{\text{SC}}(x_S, \tilde{x}_S) = JSD\left(P; \tilde{P}\right)$$
$$= \frac{1}{2}\left(D_{\text{KL}}[P||Q] + D_{\text{KL}}[\tilde{P}||Q]\right), \quad (1)$$

where $Q = (P + \tilde{P})/2$ is the average information of the original and style-diversified samples. D_{KL} denotes the KL Divergence between the posterior probability $\{P, \tilde{P}\}$ and Q. JSD is a valid statistical metric, which can be applied to any two arbitrary distributions, even with the different distribution support (Shui et al., 2022a). Thus, we can estimate and minimize the JSD for P and $\tilde{P}$ to learn the invariant semantic information across two styles.

Retrospection Consistency (RC) Backbones of generalizable models commonly start from ImageNet (Deng et al., 2009) pre-trained weights since the pre-trained backbones have learned general representation of the visual world. However, the model learns more task-specific information and fit to the source data after training, which limiting the performance on the unseen scenarios. Since ImageNet pre-trained weights are available to every generalizable model as the initialization, we propose RC to leverage such knowledge to lead the model less overfitting to the source data and to retrospect the knowledge of the visual world lying in initialization. Specifically, RC is implemented as the feature-level distance minimization between the training model θ_S and pre-trained model θ_{IN}. In addition, the style-diversified samples $\tilde{x}_S$ in style consistency is also used in RC, which can lead the generated samples close to the real-world style. Similar to style consistency, the backbone feature used for RC depends on the task. We define RC as:
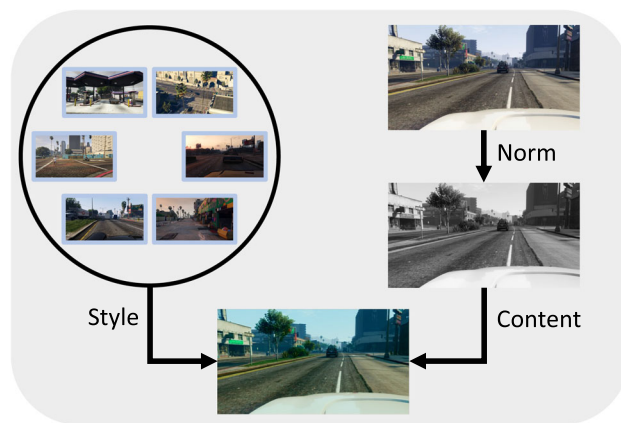

Fig. 3 Illustration of the style hallucination module

$$\mathcal{L}_{\text{RC}}(x_S, \tilde{x}_S) = (f_{\text{RC}}(x_S, \tilde{x}_S; \theta_S) - f_{\text{RC}}(x_S; \theta_{IN}))^2, \quad (2)$$

where $f_{\text{RC}}(x_S, \tilde{x}_S; \theta_S)$ denotes the bottleneck feature of original sample x_S and style-diversified sample $\tilde{x}_S$, which are obtained from the ongoing model θ_S. $f(x_S; \theta_{IN})$ denotes the bottleneck feature of original sample x_S in the retrospective ImageNet pre-trained model θ_{IN}.

Discussion Our retrospection consistency is inspired by the Feature Distance (FD) in DAFormer (Hoyer et al., 2022) but with different motivation and implementation. *First*, DAFormer focuses on unsupervised domain adaptation in semantic segmentation with unlabeled real-world data available, so it only focuses on fitting to the specific target domain (CityScapes (Cordts et al., 2016)) rather than addressing unseen domain shift. *Second*, FD in DAFormer is used to better classify those similar classes (e.g., bus and train) with the classification knowledge from ImageNet. In this paper, we focus on the domain generalization in visual recognition

tasks. Since we have no idea about the target distribution, RC in our framework serves as an important guidance for the general visual knowledge, leading the model less overfitting to the source data. In addition, apart from semantic segmentation, RC can be applied to image classification and object detection.

3.2 Applications of Dual Consistency Learning

The recognition objective is different for each task: image classification leverages the global information to classify the whole image; semantic segmentation is required to classify each pixel to the pre-defined categories; object detection aims to detect the object and recognize the class. Consequently, the implementation of dual consistency constraints is based on the task requirements.

Image Classification The input image x first forwards through the backbone θ and the global average pooling layer to obtain the backbone feature $f(x; \theta)$, which is used for retrospective consistency. Then a linear classifier is used to predict the image class for backbone features of both the original and stylized samples. The prediction after softmax is the posterior probability P and $\tilde{P}$.

Semantic Segmentation Since the pixel-level prediction is required for semantic segmentation, the consistency is applied to each pixel instead of the whole image. In addition, semantic segmentation in the urban scene contains both foreground (e.g., car and person) and background (e.g., road and wall) classes, while the ImageNet pre-trained model only focuses on the foreground categories. Thus, only the foreground backbone features are used for retrospective consistency. Here we define the mask of foreground pixels as M_{fg} and the retrospective features $f_{RC}(x; \theta)$ is $M \cdot f(x; \theta)$. Nevertheless, since the style variation is applied to both foreground and background pixels, so the style consistency is applied to the posterior probabilities of all the pixels, which are the predicted value after classifier and softmax.

Object Detection Object detection requires to classify the object category within the bounding box instead of the whole image or each pixel. Thus, the consistency should be applied only to the box features. Specifically, we use the ground truth box to obtain the RoI (Region of Interest) features from the backbone features via RoI align (He et al., 2017), and the RoI features $f_{RoI}(x)$ serve as the retrospective features $f_{RC}(x; \theta)$. After that, the RoI features are fed into the classifier to obtain the posterior probability P and $\tilde{P}$ for each object, which are used in the style consistency.

3.3 Style Hallucination

Background Style transfer (Chen et al., 2021a; Huang & Belongie, 2017) and domain generalization (Tang et al., 2021; Zhou et al., 2021a) methods show that the channel-wise

mean and standard deviation can represent the non-content style of the image, which plays an important role in the domain shift. The style features can be readily used by AdaIN (Huang & Belongie, 2017) which can transfer the image to an arbitrary style while remaining the content:

$$\text{AdaIN}(x, y) = \sigma(y) \left(\frac{x - \mu(x)}{\sigma(x)} \right) + \mu(y), \qquad (3)$$

where x and y denotes the feature maps providing the content and style respectively. $\mu(*)$ and $\sigma(*)$ denotes the channel-wise mean and standard deviation. In domain generalization, as only one or multiple source domains are accessible, previous works modify AdaIN by replacing style features with other source styles. Those styles can be directly obtained from other samples (Tang et al., 2021) or can be generated by mixing other styles with its own styles (Zhou et al., 2021a). *Style Hallucination Module (SHM)* The ways of previous methods (Tang et al., 2021; Zhou et al., 2021a) in generating extra styles are sub-optimal, since they just randomly swap or mix source styles without considering the frequency and diversity of the source styles. As a result, more samples of the dominant style (e.g., daytime) will be generated, yet the generated distribution may be quite different from the target one. Since we have no idea about the distribution of target set, it is better to diversify the source samples as much as possible. We next introduce the Style Hallucination Module (SHM) for generating diverse source samples. The illustration of SHM is shown in Fig. 3.

Definition 1 A basis B of a vector space V over a field F is a linearly independent subset of V that spans V. When the field is the reals $\mathbb{R}$, the resulting basis vectors are n-tuples of reals that span n-dimensional Euclidean space $\mathbb{R}^n$ (Halmos, 1987).

According to Definition 1, style space can be viewed as a subspace of C-dimensional vector space, and thus all possible styles can be represented by the basis vectors. However, if we directly take C linearly independent vectors as the basis, e.g., orthogonal unit vectors, many unrealistic styles are generated since the realistic styles are only in a small subspace, and such generated styles can impair the model training. To reconcile the diversity and realism, we use farthest point sampling (FPS) (Qi et al., 2017) to select C styles from all the source styles as basis styles. FPS is widely used for point cloud downsampling, which can iteratively choose C points from all the points, such that the chosen points are the most distant points with respect to the remaining points. Despite not strictly linearly independent, basis styles obtained by FPS can represent the style space to the utmost extent, and also contain many rare styles since rare styles are commonly far away from dominant ones. In addition, we recalculate the dynamic basis styles every k epochs instead
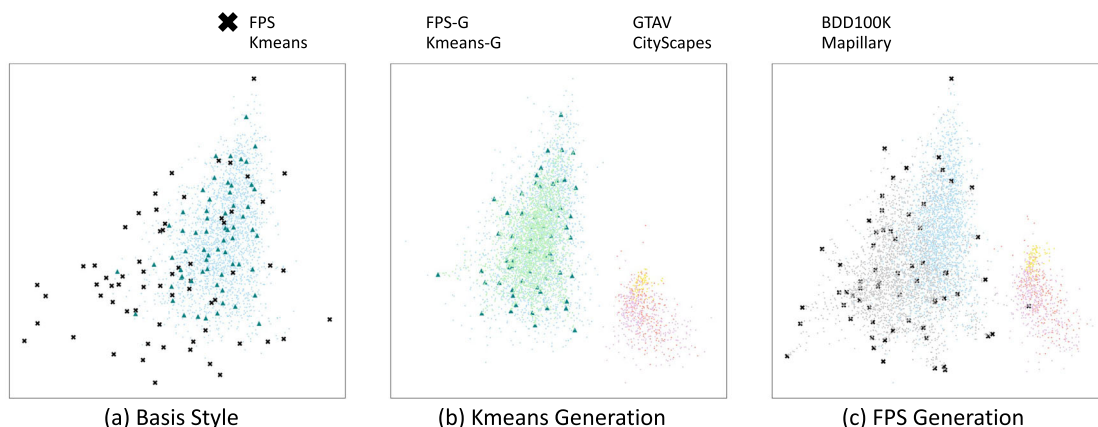
Fig. 4 Visualization of distributions of different domains. **a** Comparison of two kinds of basis styles; **b** generated style with Kmeans basis style; **c** generated style with FPS basis style (Zoom in for details)

of fixing them, as the style space is changing along with the model training. To generate new styles, we sample the combination weight $W = [w_1, \cdots, w_C]$ from Dirichlet distribution $B([\alpha_1, \cdots, \alpha_C])$ with the concentration parameters $[\alpha_1, \cdots, \alpha_C]$ all set to $1/C$. The basis styles are then linearly combined by W:

$$\mu_{HS} = W \cdot \mu_{base}, \qquad \sigma_{HS} = W \cdot \sigma_{base}, \tag{4}$$

where $\mu_{base} \in \mathbb{R}^{C \times C}$ and $\sigma_{base} \in \mathbb{R}^{C \times C}$ are the C basis styles. With the generated styles, style hallucinated samples $\tilde{x}_S$ can be obtained by:

$$\tilde{x}_S = \sigma_{HS} \left(\frac{x_S - \mu(x_S)}{\sigma(x_S)} \right) + \mu_{HS}. \tag{5}$$

Discussion We take the semantic segmentation benchmark as an example to verify the FPS basis style selection. In Fig. 4, GTAV (Richter et al., 2016) is the source domain while CityScapes (Cordts et al., 2016), BDD100K (Yu et al., 2020) and Mapillary (Neuhold et al., 2017) are the target domains. Selecting representative basis styles is crucial for SHM. FPS is adopted in our method as it can cover the rare styles to the utmost extent. Another way is taking the Kmeans (MacQueen, 1967) clustering centers as the basis. As shown in Fig. 4a, FPS samples (black cross) spread out more than Kmeans centers (teal triangle), and can cover almost all possible source styles (lightskyblue point). When using the basis styles for style generation, styles obtained from Kmeans centers (Fig. 4b) are still within the source distribution and even ignore some possible rare styles. In contrast, FPS basis styles can generate more diverse styles (Fig. 4c), and even generate some styles close to the real-world ones (yellow, pink and orange point). Table 8 further demonstrates the effectiveness of FPS basis styles and shows that the Kmeans basis styles

are even worse than directly swapping and mixing source styles.

3.4 Training Objective

The overall training objective is the combination of the task loss and the proposed two consistency constraints:

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\text{task}}(x_S, y_S) &+ \lambda_{\text{SC}} \mathcal{L}_{\text{SC}}(x_S, \tilde{x}_S) \\ &+ \lambda_{\text{RC}} \mathcal{L}_{\text{RC}}(x_S, \tilde{x}_S), \end{aligned} \tag{6}$$

where λ_{SC} and λ_{RC} are the weights for style consistency and retrospection consistency, respectively. The task loss is the cross entropy loss for image classification and semantic segmentation, and is the RPN loss together with cross entropy loss for object detection.

4 Experiments

In this section, we conduct experiments on image classification, semantic segmentation and object detection to demonstrate the superiority of SHADE. We first compare our method with state-of-the-art methods in the three visual tasks. Then, we provide ablation studies on image classification and semantic segmentation to verify the effectiveness of each proposed component. Finally, we analyze and evaluate SHADE on semantic segmentation and provide the visualization results.

4.1 Semantic Segmentation

4.1.1 Experimental Setup

Datasets Two synthetic datasets (GTAV (Richter et al., 2016) and SYNTHIA (Ros et al., 2016)) and three real-world

datasets (CityScapes (Cordts et al., 2016), BDD100K (Yu et al., 2020) and Mapillary (Neuhold et al., 2017)) are used in this paper. GTAV (Richter et al., 2016) contains 24,966 images with the size of 1914×1052, splitting into 12,403 training, 6382 validation, and 6181 testing images. SYN-THIA (Ros et al., 2016) contains 9400 images of 960×720, where 6,580 images are used for training and 2820 images for validation. CityScapes (Cordts et al., 2016) contains 2975 training images and 500 validation images of 2048×1024. BDD100K (Yu et al., 2020) and Mapillary (Neuhold et al., 2017) contain 7000 and 18,000 images for training, and 1000 and 2000 images for validation, respectively.

Implementation Details We use both ConvNets and Transformer backbone to conduct experiments. For the *ConvNets models*, we use DeepLabV3+ (Chen et al., 2018) as the segmentation model and equip it with two backbones: ResNet-50 and ResNet-101 (He et al., 2016). The SHM is inserted after the first Conv-BN-ReLU layer (layer0). We re-select the basis styles with the interval $k = 3$. We set $\lambda_{SC} = 10$ and $\lambda_{RC} = 1$. Models are optimized by the SGD optimizer with the learning rate 0.01, momentum 0.9 and weight decay 5×10^{-4}. The polynomial decay (Liu et al., 2015) with the power of 0.9 is used as the learning rate scheduler. All models are trained with the batch size of 8 for 40K iterations. During training, four widely used data augmentation techniques are adopted, including color jittering, Gaussian blur, random flipping, and random cropping of 768×768. For the *Transformer models*, we use SegFormer (Xie et al., 2021) with MiT-B5 (Xie et al., 2021) as the segmentation model. SHM is inserted after the first block and the basis styles are re-selected every 4k iterations. We set $\lambda_{SC} = 10$ and $\lambda_{RC} = 0.005$. Following Hoyer et al. (2022), models are optimized by the AdamW (Loshchilov & Hutter, 2019) optimizer with the learning rate 6×10^{-5} for encoder and 6×10^{-4} for decoder, and weight decay 0.01. Linear learning rate warmup by 1.5k iterations is first adopted, and then the learning rate linearly decays. All models are trained with the batch size of 2 for 40K iterations.

Protocols In this section, we focus on three protocols. (1) *Synthetic-to-real single-source DG* takes the GTAV as the source domain and evaluates the model on the three real-world target domains. To conduct a fair comparison with Choi et al. (2021) and Huang et al. (2021), we train the model with GTAV training data (12,403 images) when using ResNet-50 backbone (same with Choi et al. (2021)), and with the whole GTAV datasets (24,966 images) when using ResNet-101 backbone (same with Huang et al. (2021)) and MiT-B5 backbone. (2) *Synthetic-to-real multi-source DG* takes the GTAV and SYNTHIA as the source domains and evaluate on the three real-world target domains. We follow Choi et al. (2021) to train the model with the training set of GTAV (12,403 images) and SYNTHIA (6580 images) using ResNet-50 backbone. (3) *Real-to-others single-source DG* is

also investigated in this paper. Following Choi et al. (2021), we train the model with the training set of CityScapes and evaluate it on the other four synthetic and real-world datasets. *Evaluation Metric* We use the 19 shared semantic categories for training and evaluation. The mean intersection-over-union (mIoU) of the 19 categories on the target datasets is adopted as the evaluation metric.

4.1.2 Comparison with State-of-the-Art Methods

Synthetic-to-Real Single-Source DG In Table 1, we compare SHADE with state-of-the-art methods under single-source setting, including SW (Pan et al., 2019), IterNorm (Huang et al., 2019), IBN-Net (Pan et al., 2018), ISW (Choi et al., 2021), DRPC (Yue et al., 2019), FSDR (Huang et al., 2021), GTR (Peng et al., 2021), SAN-SAW (Peng et al., 2022) and AdvStyle (Zhong et al., 2022). *First*, we compare models that are trained with GTAV training set, using ResNet-50 backbone. SHADE achieves an average mIoU of 42.42% on the three real-world target datasets, yielding an improvement of 15.00% mIoU over the baseline and outperforming the previous best method (AdvStyle) by 2.39%. *Second,* we further compare methods under the training protocol of DRPC (Yue et al., 2019) and FSDR (Huang et al., 2021), using ResNet-101 backbone and taking the whole set of GTAV (24,966 images) as the training data. Note that DRPC (Yue et al., 2019) and FSDR (Huang et al., 2021) utilize extra real-world data from ImageNet (Deng et al., 2009) or even driving scenes. Moreover, they select the best checkpoint for each target dataset, which is impractical in the real-world applications. Even so, we achieve the best results on all three datasets, 46.66% on CityScapes, 43.66% on BDD100K, and 45.50% on Mapillary, outperforming FSDR (Huang et al., 2021) and AdvStyle (Zhong et al., 2022) by by 2.14% and 2.85% in the average mIoU. *Third*, we further investigate SHADE with the Transformer backbone. Since no previous work leveraging the Transformer backbone, we reproduce MixStyle (Zhou et al., 2021a), CrossNorm (Tang et al., 2021) and AdvStyle (Zhong et al., 2022) based on the source code. As shown in Table 1, the source only model serves as a strong baseline and these method can hardly improve the performance. Instead, SHADE outperforms the baseline by 7.67% on CityScapes, 3.33% on BDD100K, and 6.27% on Mapillary, respectively. These results show that we produce new state of the art in domain generalized semantic segmentation with different segmentation models and backbones.

Synthetic-to-Real Multi-Source DG To further verify the effectiveness of SHADE, we compare SHADE with IBN-Net (Pan et al., 2018), ISW (Choi et al., 2021), PTM (Kim et al., 2022) and AdvStyle (Zhong et al., 2022) under the multi-source setting. We use ResNet-50 as the backbone and take the training set of GTAV and SYNTHIA as the source domains. As shown in Table 2, SHADE gains an improve-

Table 1 Comparison with state-of-the-art methods on single-source DG-Seg with ResNet-50, ResNet-101, and MiT-B5 backbones, respectively

Net	Methods (GTAV)	Venue	Extra data	CityScapes	BDD100K	Mapillary	Avg.
ResNet-50	Baseline	–	×	28.95	25.14	28.18	27.42
	IBN-Net (Pan et al., 2018)	ECCV 18	×	33.85	32.30	37.75	34.63
	IterNorm (Huang et al., 2019)	CVPR 19	×	31.81	32.70	33.88	32.79
	SW (Pan et al., 2019)	ICCV 19	×	29.91	27.48	29.71	29.03
	DRPC (Yue et al., 2019)[§]	ICCV 19	✓	37.42	32.14	34.12	34.56
	ISW (Choi et al., 2021)	CVPR 21	×	36.58	35.20	40.33	37.37
	GTR (Peng et al., 2021)	TIP 21	×	37.53	33.75	34.52	35.27
	SAN-SAW (Peng et al., 2022)	CVPR 22	×	39.75	37.34	41.86	39.65
	AdvStyle (Zhong et al., 2022)	NeurIPS 22	×	39.60	38.59	41.89	40.03
	Ours	–	×	**44.65**	**39.28**	**43.34**	**42.42**
ResNet-101	Baseline	–	×	32.97	30.77	30.68	31.47
	IBN-Net (Pan et al., 2018)	ECCV 18	×	37.37	34.21	36.81	36.13
	DRPC (Yue et al., 2019)[§]	ICCV 19	✓	42.53	38.72	38.05	39.77
	ISW (Choi et al., 2021)[†]	CVPR 21	×	37.20	33.36	35.57	35.38
	FSDR (Huang et al., 2021)[§]	CVPR 21	✓	44.80	41.20	43.40	43.13
	GTR (Peng et al., 2021)	TIP 21	×	43.70	39.60	39.10	40.80
	SAN-SAW (Peng et al., 2022)	CVPR 22	×	45.33	41.18	40.77	42.43
	AdvStyle (Zhong et al., 2022)	NeurIPS 22	×	44.51	39.27	43.48	42.42
	Ours	–	×	**46.66**	**43.66**	**45.50**	**45.27**
MiT-B5	Baseline	–	×	45.60	44.86	48.72	46.39
	MixStyle (Zhou et al., 2021a)[†]	ICLR 21	×	43.29	44.23	46.51	44.68
	CrossNorm (Tang et al., 2021)[†]	ICCV 21	×	46.41	44.69	50.21	47.10
	AdvStyle (Zhong et al., 2022)[†]	NeurIPS 22	×	46.56	45.10	48.35	46.67
	Ours	–	×	**53.27**	**48.19**	**54.99**	**52.15**

"Extra Data" denotes using extra real-world data during training. § denotes selecting best checkpoint for each target dataset, and † denotes reproducing the results based on the source code
Bold values denote the best results

Table 2 Comparison with state-of-the-art methods on multi-source DG-Seg

Methods (G+S)	Venue	CityScapes	BDD100K	Mapillary	Avg.
Baseline	–	35.46	25.09	31.94	30.83
IBN-Net (Pan et al., 2018)	ECCV 18	35.55	32.18	38.09	35.27
MLDG (Li et al., 2018a)	AAAI 18	38.84	31.95	35.60	35.46
ISW (Choi et al., 2021)	CVPR 21	37.69	34.09	38.49	36.76
PTM (Kim et al., 2022)	CVPR 22	44.51	38.07	42.70	41.76
AdvStyle (Zhong et al., 2022)	NeurIPS 22	39.29	39.26	41.14	39.90
Ours	–	**47.43**	**40.30**	**47.60**	**45.11**

All models use ResNet-50 backbone and are trained with training sets of GTAV and SYNTHIA
Bold values denote the best results

ment of 14.28% in average mIoU over the baseline, and outperforms PTM (Kim et al., 2022) and AdvStyle (Zhong et al., 2022) by 3.35% and 5.21% respectively. The significant improvement over previous works is mainly benefited from the various samples. With richer source samples, our SHM can generate more informative and diverse styles, which can effectively facilitate the dual consistency learning.

Real-to-Others Single-Source DG In Table 3, we compare SHADE with previous methods under the real-to-others DG setting, where CityScapes (Cordts et al., 2016) is leveraged as the source domain and the model is generalized to real (BDD100K (Yu et al., 2020) and Mapillary (Neuhold et al., 2017)) and synthetic (GTAV (Richter et al., 2016) and SYN-THIA (Ros et al., 2016)) domains. As shown in Table 3, SHADE consistently outperforms ISW (Choi et al., 2021)

Table 3 Comparison with state-of-the-art methods on Real-to-Others DG-Seg

Methods	GTAV	SYNTHIA	BDD100K	Mapillary	Avg.
Baseline	42.55	23.29	44.96	51.68	40.62
SW (Pan et al., 2019)	44.87	26.10	48.49	55.82	43.82
IterNorm (Huang et al., 2019)	45.73	25.98	49.23	56.26	44.30
IBN-Net (Pan et al., 2018)	45.06	26.14	48.56	57.04	44.20
ISW (Choi et al., 2021)	45.00	26.20	50.73	58.64	45.14
Ours	**48.61**	**27.62**	**50.95**	**60.67**	**46.96**

All models use ResNet-50 backbone and are trained with GTAV training set

Bold values denote the best results

Table 4 Comparison with state-of-the-art methods on PACS image classification benchmark

Method	Art	Car	Ske	Pho	Avg.
ERM	67.4	74.4	51.4	42.6	58.9
JiGen (Carlucci et al., 2019)	69.1	74.6	52.4	41.5	59.4
RSC (Huang et al., 2020)	68.8	74.5	53.6	41.9	59.7
L2D (Wang et al., 2021c)	74.3	77.5	54.4	45.9	63.0
ERM+SHADE	**76.5**	77.2	**60.1**	49.3	65.8
RSC+SHADE	75.4	74.8	55.8	49.8	64.0
L2D+SHADE	76.1	**79.7**	**60.1**	**51.0**	**66.7**

One domain (name in column) is used as the source (training) data and the other domains are used as the target (testing) data

Bold values denote the best results

and IBN-Net (Pan et al., 2018) on both real and synthetic datasets. These results further verify the versatility of our method.

4.2 Image Classification

4.2.1 Experimental Setup

Datasets We conduct experiments on the PACS (Li et al., 2017) benchmark. PACS (Li et al., 2017) contains four domains (Artpaint, Cartoon, Sketch, and Photo), and each domain contains 224×224 images belonging to seven categories. There are 9991 images in total.

Implementation Details Following Huang et al. (2020), we use the ResNet18 (He et al., 2016) pretrained on ImageNet (Deng et al., 2009) as the backbone. The SHM is inserted after the first Conv-BN-ReLU layer and the basis styles are selected every 3 epochs. The loss weight λ_{SC} and λ_{RC} are set to 10 and 0.1, respectively. We train the model by SGD optimizer. The learning rate is initially set to 0.004 and divided by 10 after 24 epochs. The model is trained for 30 epochs in total with a batch size of 128.

Evaluation Metric We select one of the four domains in PACS as the source domain and the other domains as the target domains. The mean accuracy of the 7 categories on the target datasets is adopted as the evaluation metric.

4.2.2 Comparison with State-of-the-Art Methods

We compare SHADE with the baseline (ERM (Vapnik, 2013)) and three state-of-the-art DG methods, including JiGen (Carlucci et al., 2019), RSC (Huang et al., 2020) and L2D (Wang et al., 2021c). We reproduce JiGen (Carlucci et al., 2019), RSC (Huang et al., 2020) and L2D (Wang et al., 2021c) with their official source codes. All methods use the same baseline (ERM (Vapnik, 2013)). To verify the effectiveness of SHADE, we apply SHADE to ERM (Vapnik, 2013), RSC (Huang et al., 2020) and L2D (Wang et al., 2021c). As shown in Table 4, combining SHADE with ERM (Vapnik, 2013) can outperform state-of-the-art method (L2D (Wang et al., 2021c)) on all four domains by 2.2%, 0.3%, 5.7%, and 3.4%, respectively. In addition, SHADE can yield improvements on all the three models by 6.9% on ERM (Vapnik, 2013), 4.3% on RSC (Huang et al., 2020), and 3.7% on L2D (Wang et al., 2021c) in the average accuracy. The above results demonstrate SHADE is versatile and effective.

4.3 Object Detection

4.3.1 Experimental Setup

Datasets We use the urban-scene detection benchmark (Wu & Deng, 2022) in this paper. The benchmark includes five datasets of different weather conditions. Specifically, daytime-sunny dataset contains 19,395 and 8313 images from BDD100K (Yu et al., 2020) for training and testing, respectively. The night-sunny dataset contains 26,158 images from BDD100K. The dusk-rainy and night-rainy datasets include 3501 and 2494 images rendered from BDD100K by Wu et al. (2021). In addition, 3775 foggy images from FoggyCityscapes (Sakaridis et al., 2018) and Adverse-Weather (Hassaballah et al., 2020) form the daytime-foggy dataset.

Implementation Details Following Wu and Deng (2022), we use Faster R-CNN (Ren et al., 2015) with ResNet-101 (He et al., 2016) backbone as the detector. The SHM is inserted after the first Conv-BN-ReLU layer and the basis styles are selected every epoch. The loss weight λ_{SC} and λ_{RC} are set to

Table 5 Comparison with state-of-the-art methods on domain generalized object detection benchmark

Method	Night Sunny	Dusk Rainy	Night Rainy	Day Foggy	Avg.
Baseline	33.5	26.6	14.5	31.9	26.6
SW (Pan et al., 2019)	33.4	26.3	13.7	30.8	26.1
IBN-Net (Pan et al., 2018)	32.1	26.1	14.3	29.6	25.5
IterNorm (Huang et al., 2019)	29.6	22.8	12.6	28.4	23.4
ISW (Choi et al., 2021)	33.2	25.9	14.1	31.8	26.3
Single-DGOD (Wu & Deng, 2022)	**36.6**	28.2	16.6	**33.5**	**28.7**
Ours	33.9	**29.5**	**16.8**	33.4	28.4

Bold values denote the best results

Table 6 Ablation studies of each component on the synthetic-to-real single source DG benchmark

No	SHM	$\mathcal{L}_{SC}$	$\mathcal{L}_{RC}$	EMA	CityScapes	BDD100K	Mapillary	Avg.
1	×	×	×	×	28.95	25.14	28.18	27.42
2	✓	×	×	×	38.68	32.40	35.96	35.68
3	✓	✓	×	×	42.66	35.92	40.42	39.67
4	✓	×	✓	×	41.43	37.65	41.77	40.29
5	✓	✓	×	✓	42.38	38.04	42.34	40.92
6	✓	✓	✓	×	**44.65**	**39.28**	**43.34**	**42.42**

All models use ResNet-50 backbone and are trained with GTAV training set. SHM: style hallucination module; EMA: using exponential moving average model instead of ImageNet pre-trained model

Bold values denote the best results

10 and 0.1, respectively. SGD optimizer is used to train the model with learning rate 0.004, momentum 0.9 and weight decay 5×10^{-4}. The model is trained for 10 epochs in total with a batch size of 2.

Evaluation Metric The daytime-sunny dataset is used as the source training data, while the night-sunny, dusk-rainy, night-rainy and daytime-foggy datasets are used for the target testing data. We use the 7 shared categories for training and evaluation and the mean average precision (mAP) with a threshold of 0.5 is adopted as the evaluation metric.

4.3.2 Comparison with State-of-the-art Methods

We compare SHADE with the baseline model, SW (Pan et al., 2019), IBN-Net (Pan et al., 2018), IterNorm (Huang et al., 2019), ISW (Choi et al., 2021) and Single-DGOD (Wu & Deng, 2022) in Table 5. The results are borrowed from Wu and Deng (2022). As shown in Table 5, SHADE outperforms the baseline model on all the target datasets by 0.4% on night-sunny, 2.9% on dusk-rainy, 2.3% on night-rainy, and 1.5% on daytime-foggy. Furthermore, SHADE achieves the state-of-the-art performance on dusk-rainy and night-rainy datasets. These results demonstrate that SHADE is applicable to improving the generalization ability for the object detection model.

Table 7 Ablation studies of each component on the PACS benchmark

No	SHM	$\mathcal{L}_{SC}$	$\mathcal{L}_{RC}$	Art	Car	Ske	Pho	Avg.
1	×	×	×	67.4	74.4	51.4	42.6	58.9
2	✓	×	×	69.9	73.9	51.1	43.4	59.6
3	✓	✓	×	72.8	74.8	60.0	49.1	64.2
4	✓	×	✓	69.6	75.8	54.3	43.5	60.8
5	✓	✓	✓	**76.5**	**77.2**	**60.1**	**49.3**	**65.8**

ERM baseline with ResNet-18 backbone is used. SHM: style hallucination module

Bold values denote the best results

4.4 Ablation Studies

To investigate the effectiveness of each component in SHADE, we conduct ablation studies in Tables 6 and 7 on semantic segmentation and image classification, respectively.

Effectiveness of Style Hallucination Module (SHM) SHM is the basis of SHADE. When using SHM only, we directly apply cross entropy loss on the style hallucinated samples. As shown in the second row of Table 6, our SHM can largely improves the model performance even without using the proposed dual consistency learning. This demonstrates the importance of training the model with diverse samples and the effectiveness of our SHM.

Effectiveness of Style Consistency (SC) SC is the consistency constraint that leads the model to learn style invariant repre-

Table 8 Comparison of different style variation methods

Methods (GTAV)	CityScapes	BDD100K	Mapillary	Avg.
Baseline	28.95	25.14	28.18	27.42
Random Style	37.99	37.63	38.06	37.89
MixStyle (Zhou et al., 2021a)	43.14	37.94	42.22	41.10
CrossNorm (Tang et al., 2021)	43.13	37.20	41.83	40.72
Kmeans Basis	40.50	37.62	39.46	39.19
Ours	**44.65**	**39.28**	**43.34**	**42.42**

All models use ResNet-50 backbone and are trained with GTAV training set
Bold values denote the best results

(a) Location of SHM.

(b) Basis style selection interval.
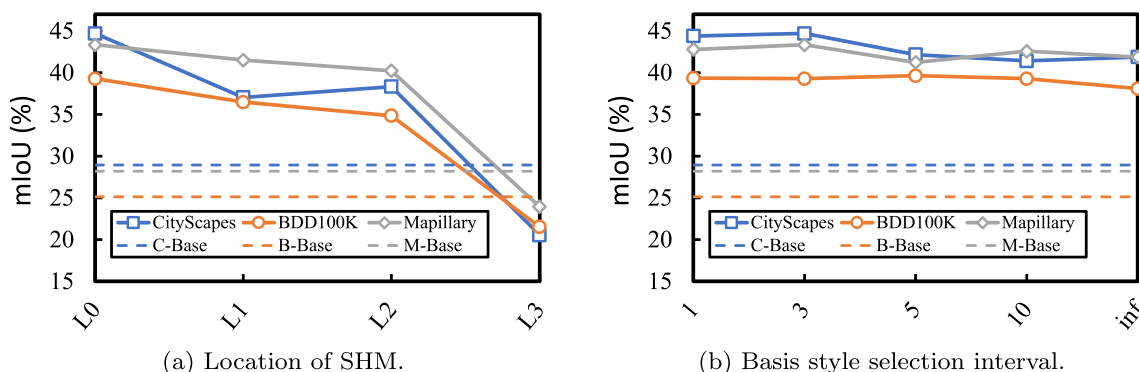
Fig. 5 Parameter analysis. "C-Base", "B-Base" and "M-Base" denote the performance of baseline model on CityScapes, BDD100K and Mapillary, respectively

sentation. In Table 6, compared with only applying SHM, SC yields an improvement of 3.99% in average mIoU, demonstrating the superiority of the proposed logit pairing over cross entropy loss in learning style invariant model.

Effectiveness of Retrospection Consistency (RC) *First*, RC serves as an important guidance to ameliorate the overfitting problem by general visual knowledge. Applying RC on top of SHM can yield an improvement of 4.61%, while removing RC will degrade the performance of SHADE by 2.75% in mIoU. *Second*, we conduct experiments to verify that the effectiveness of RC lies in the general visual knowledge instead of feature-level distance minimization of the paired samples. As directly minimizing the feature-level absolute distance of paired samples will lead to sub-optimal results (lead all the features close to zero), we replace the ImageNet pre-trained model in RC by exponential moving average (EMA) model. Comparing the fifth row and the sixth row in Table 6, EMA model only gains 1.25% improvement while RC improves the SC model by 2.75%. The results verify the significance of the retrospective knowledge in RC.

Ablation Studies on Image Classification We further conduct ablation studies on the PACS benchmark to verify the effectiveness of each component in SHADE. Different phenomena from the segmentation benchmark are observed. *First*, adopting SHM can only slightly improve the model performance since the styles of a single classification dataset are limited. *Second*, the performance yields a significant

improvement of 4.6% in average accuracy when style consistency is applied. *Third*, leveraging retrospective consistency can gain an improvement of 1.2% and applying it on top of the style consistency can further improve the performance by 1.6% in average accuracy.

4.5 Further Evaluation

In this section, we compare different style variation methods under the proposed dual consistency constraints and analyze two important hyper-parameter, i.e., the location of SHM and the basis style selection interval, in our framework. Experiments in this section are based on the synthetic-to-real single-source DG in semantic segmentation. ResNet-50 (He et al., 2016) is leveraged as the backbone.

Comparison of different style variation methods We compare SHM with random style, MixStyle (Zhou et al., 2021a), CrossNorm (Tang et al., 2021) and style hallucination with Kmeans basis in Table 8. Random style utilizes the sampled styles from the standard normal distribution to form new samples. MixStyle (Zhou et al., 2021a) generates new styles by mixing the original style with the randomly shuffled style within a mini-batch, and CrossNorm (Tang et al., 2021) swaps the original style with another style within the shuffled mini-batch. SHM and Kmeans basis both use the linear combination of basis style to generate new styles, but the basis styles of SHM are selected by FPS (Qi et al.,

Table 9 Analysis of loss form of style consistency

Methods (GTAV)	CityScapes	BDD100K	Mapillary	Avg.
SHADE w.o. SC	41.43	37.65	41.77	40.29
SHADE-MMD	42.51	37.33	42.30	40.71
SHADE-WD	43.25	37.71	41.87	40.94
SHADE	**44.65**	**39.28**	**43.34**	**42.42**

Bold values denote the best results

Table 10 Analysis of loss form of retrospection consistency

Methods (GTAV)	CityScapes	BDD100K	Mapillary	Avg.
SHADE w.o. RC	42.66	35.92	40.42	39.67
SHADE-L1	43.39	37.98	43.16	41.51
SHADE	**44.65**	**39.28**	**43.34**	**42.42**

Bold values denote the best results

2017) while those of Kmeans basis are obtained by Kmeans clustering centers (MacQueen, 1967). We can make four observations from Table 8. *First*, we cannot make full use of dual consistency to achieve significant performance with the unrealistic random styles since standard normal distribution cannot represent the source nor the target domains. *Second*, despite the use of realistic source styles, random utilization of MixStyle and CrossNorm leads to the generation of more samples from the dominant styles that may be different from the target styles. When using MixStyle and CrossNorm, the model achieves an average mIoU of 41.10% and 40.72%, respectively. *Third*, as shown in Fig. 4b, Kmeans basis suffers from the similar but more severe dominant style issue in style generation. As a result, rare styles are almost discarded and thus the model achieves poorer performance than the above two. *Fourth*, SHM selects basis styles via FPS, and thus the selection can cover the source distribution to a large extent, especially those rare styles. With such basis styles, SHM generates styles from all the source distributions, and some generated styles are even close to the target domains (Fig. 4c). Consequently, combining SHM with dual consistency learning, SHADE can reap the benefit of the source data and outperform other methods on all three target datasets.

Location of SHM We investigate the impact of inserting SHM in different locations in Fig. 5a. "L0" denotes inserting SHM after the first Conv-BN-ReLU layer (layer0) and "L1" to "L3" denote inserting SHM after the corresponding (1–3) ResNet layer. As shown in Fig. 5a, "L0" achieves the best result while the performance of "L1" and "L2" drops a little. However, the model suffers from drastic performance degradation when inserting SHM after layer3, and the performance is even worse than the baseline. The reasons are two-fold. *First*, the channel-wise mean and standard deviation represent more style information in the shallow layers of deep neural networks while they contain more semantic informa-

tion in deep layers (Huang & Belongie, 2017; Dumoulin et al., 2017). *Second*, the residual connections in ResNet will lead the ResNet activations of deep layers to have large peaks and small entropy, which makes the style features biased to a few dominant patterns instead of the global style (Wang et al., 2021b). The bias problem might be formulated and then alleviated by a bilevel objective (Shui et al., 2022c), which deserves further investigation. Based on the above observations, we insert SHM after the shallow layer of the backbone, i.e., layer0 of the ResNet (He et al., 2016) and block1 of MiT-B5 (Xie et al., 2021).

Basis Style Selection Interval The distribution of source styles is varied along with the model training. To better represent the style space, we re-select the basis styles with the interval of k epochs. The abscissa of Fig. 5b denotes the selection interval k and "inf" denotes only selecting the basis style once in the beginning of training. As shown in Fig. 5b, the model achieves consistent and good performance with frequent re-selection ($k <= 3$) while the performance degrades with the increase of selection interval, and the average mIoU is lower than 41% when only selecting once. Taking both the performance and computational cost into consideration, we set $k = 3$ in DG-Seg with ResNet backbone. *Note that* the batch size and training data are different across the models and tasks, and thus we follow the frequent re-selection observation to set the frequency of DG-Seg with MiT-B5 backbone and DG-Det to 4k iterations and 1 epoch, respectively.

Analysis of Loss Form of Style Consistency Jensen-Shannon Divergence (JSD) is leveraged in our framework to align the distributions of the original image and stylized image, which can help to learn the invariant semantic information across styles. In addition to JSD, Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) and Wasserstein Distance (WD) (Arjovsky et al., 2017) are also widely adopted to measure distributions. Therefore, we compare JSD with them in Table 9. *First*, compared with the model without style consistency, all three loss forms improve the generalization ability to unseen domains, further verifying the effectiveness of our framework. *Second*, JSD achieves the best performance among the three loss forms, demonstrating that JSD is better at aligning the style distributions in urban scenes.

Analysis of Loss Form of Retrospection Consistency Retrospection consistency aims to acquire general visual knowledge from the pre-trained model and to prevent the model from overfitting to the source data. In this paper, we leverage MSE loss to calculate the loss. As shown in Table 10, MSE loss achieves better performance than L1 loss in all three datasets. Moreover, both losses outperform the model without retrospection consistency by a large margin, demonstrating the superiority of our framework.
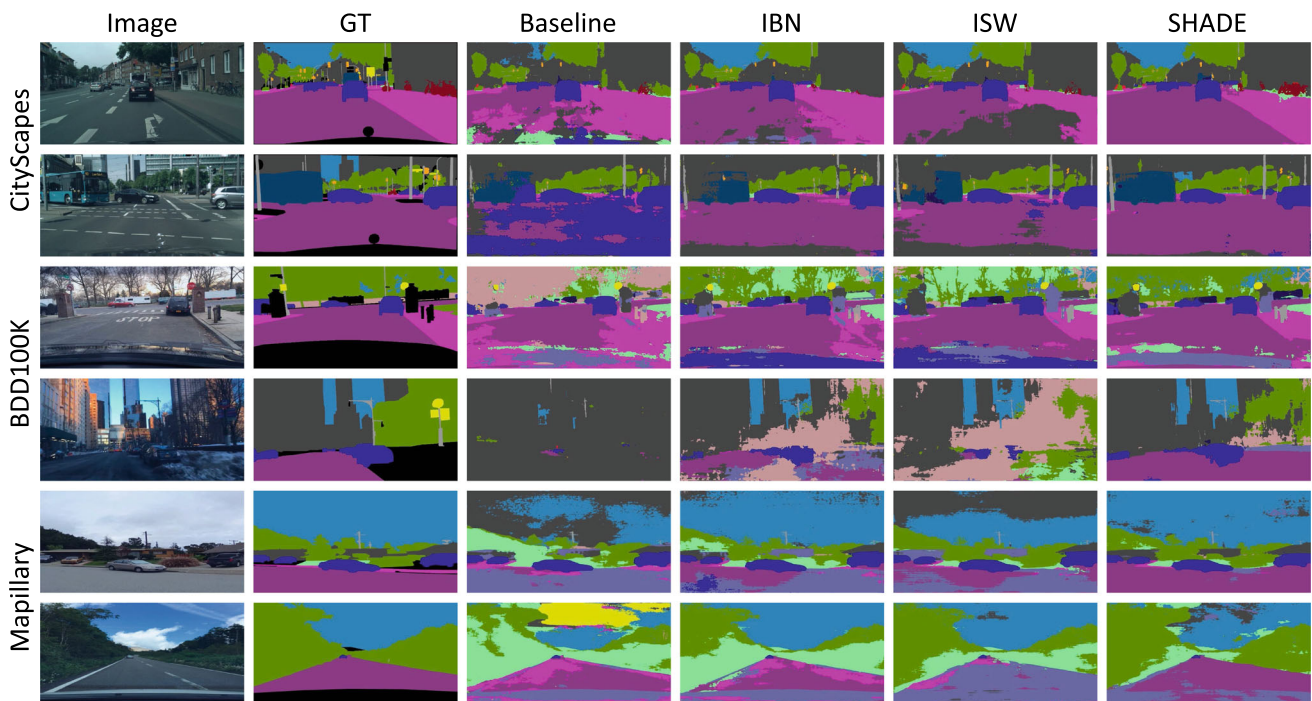
Fig. 6 Visualization of style-diversified samples

Fig. 7 Qualitative comparison of segmentation results

4.6 Visualization

Style Visualization To better understand our SHM, we visualize the style-diversified samples with an auto-encoder. We take the fixed pre-trained blocks before SHM as the encoder, and add an additional decoder. The decoder is trained by L1 loss with the source data. After training, we use SHM to generate multiple styles and replace the style of the original sample. Then we use decoder to generate the visualization results. Examples of the three tasks are shown in Fig. 6. SHM replaces the original style features with the combination of

basis styles to obtain new samples of different styles, e.g., weather change and time change.

Qualitative results To demonstrate the effectiveness of SHADE, we compare the qualitative results of semantic segmentation and object detection. We compare the segmentation results among baseline, IBN-Net (Pan et al., 2018), ISW (Choi et al., 2021) and SHADE on CityScapes (Cordts et al., 2016), BDD100K (Yu et al., 2020) and Mapillary (Neuhold et al., 2017) in Fig. 7. We obtain two observations from Fig. 7. *First*, SHADE consistently outperforms other methods under different target conditions (e.g., sunny, cloudy and overcast). *Second*, SHADE can well deal with

Fig. 8 Qualitative comparison of detection results

both background classes (e.g., road) and foreground classes (e.g., bus and bicycle). In addition, we compare SHADE with the baseline model on object detection benchmark in Fig. 8 and SHADE consistently outperforms the baseline under different environmental conditions. The above observations demonstrate that SHADE is robust to style variation and has strong ability in addressing unseen images.

5 Conclusion

In this paper, we present a novel framework (SHADE) for visual domain generalization. To address the distribution shift between the source and unseen target domains, SHADE leverages two consistency constraints to learn the domain-invariant representation by seeking consistent representation across styles and the guidance of retrospective knowledge. In addition, the style hallucination module (SHM) is equipped into our framework, which can effectively catalyze the dual consistency learning by generating diverse and realistic source samples. The proposed SHADE is effective and versatile, which can be applied to image classification, semantic segmentation and object detection tasks with both ConvNets and Transformer backbone, and can achieve state-of-the-art performance on different benchmarks and under different settings.

Data Avaiibility The datasets analysed during the current study are available as follows: GTAV (Richter et al., 2016): https://download. visinf.tu-darmstadt.de/data/from_games/ SYNTHIA (Ros et al., 2016): http://synthia-dataset.net/ CityScapes (Cordts et al., 2016): https:// www.cityscapes-dataset.com/ BDD100K (Yu et al., 2020): https://
bdd-data.berkeley.edu/ Mapillary (Neuhold et al., 2017): https://www.mapillary.com/dataset/vistas PACS (Li et al., 2017): https://domaingeneralization.github.io/\#data Urban-scene Detection (Wu & Deng, 2022): https://github.com/AmingWu/Single-DGOD

Code Availability The code is available at https://github.com/HeliosZhao/SHADE-VisualDG
.

References

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. In *ICML*

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *ECCV*

Carlucci, F. M., D'Innocente, A., Bucci, S., Caputo, B., & Tommasi, T. (2019). Domain generalization by solving jigsaw puzzles. In *CVPR*

Chen, H., Zhao, L., Zhang, H., Wang, Z., Zuo, Z., Li, A., Xing, W., & Lu, D. (2021a). Diverse image style transfer via invertible cross-space mapping. In *ICCV*

Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*

Chen, M., Zheng, Z., Yang, Y., & Chua, T. S. (2022). PiPa: Pixel-and patch-wise self-supervised learning for domain adaptative semantic segmentation. arXiv preprint arXiv:2211.07609

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *ICML*

Chen, Y., Wang, H., Li, W., Sakaridis, C., Dai, D., & Van Gool, L. (2021). Scale-aware domain adaptive faster R-CNN. *IJCV, 129*, 2223–2243.

Choi, S., Jung, S., Yun, H., Kim, J. T., Kim, S., & Choo, J. (2021). RobustNet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *CVPR*

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *CVPR*

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *CVPR*

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., & Uszkoreit, J. (2021). An image is worth 16 x 16 words: Transformers for image recognition at scale. In *ICLR*

Du, D., Chen, J., Li, Y., Ma, K., Wu, G., Zheng, Y., & Wang, L. (2022). Cross-domain gated learning for domain generalization. *IJCV, 130*, 2842–2857.

Dumoulin, V., Shlens, J., & Kudlur, M. (2017). A learned representation for artistic style. In *ICLR*

Fini, E., Sangineto, E., Lathuilière, S., Zhong, Z., Nabi, M., & Ricci, E. (2021). A unified objective for novel class discovery. In *ICCV*

French, G., Laine, S., Aila, T., Mackiewicz, M., & Finlayson, G. (2020). Semi-supervised semantic segmentation needs strong, varied perturbations. In *BMVC*

Gong, R., Li, W., Chen, Y., Dai, D., & Van Gool, L. (2021). DLOW: Domain flow and applications. *IJCV, 129*, 2865–2888.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). *A kernel two-sample test*. JMLR.

Halmos, P. R. (1987). *Finite-dimensional vector spaces*. Springer.

Hassaballah, M., Kenk, M. A., Muhammad, K., & Minaee, S. (2020). Vehicle detection and tracking in adverse weather using a deep

learning framework. *IEEE Transactions on Intelligent Transportation Systems, 22*, 4230–4242.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In *ICCV*

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *CVPR*

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., & Lakshminarayanan, B. (2020). AugMix: A simple data processing method to improve robustness and uncertainty. In *ICLR*

Hoffman, J., Wang, D., Yu, F., & Darrell, T. (2016). FCNs in the wild: Pixel-level adversarial and constraint-based adaptation. arXiv preprint arXiv:1612.02649

Hoyer, L., Dai, D., & Van Gool, L. (2022). DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *CVPR*

Huang, J., Guan, D., Xiao, A., & Lu, S. (2021). FSDR: Frequency space domain randomization for domain generalization. In *CVPR*

Huang, L., Zhou, Y., Zhu, F., Liu, L., & Shao, L. (2019). Iterative normalization: Beyond standardization towards efficient whitening. In *CVPR*

Huang, X., & Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*

Huang, Z., Wang, H., Xing, E. P., & Huang, D. (2020). Self-challenging improves cross-domain generalization. In *ECCV*

Kannan, H., Kurakin, A., & Goodfellow, I. (2018). Adversarial logit pairing. In *ICML*

Kim, J., Lee, J., Park, J., Min, D., & Sohn, K. (2022). Pin the memory: Learning to generalize semantic segmentation. In *CVPR*

Lee, S., Seong, H., Lee, S., & Kim, E. (2022). WildNet: Learning domain generalized semantic segmentation from the wild. In *CVPR*

Li, D., Yang, Y., Song, Y. Z., & Hospedales, T. M. (2017). Deeper, broader and artier domain generalization. In *ICCV*

Li, D., Yang, Y., Song, Y. Z., & Hospedales, T. (2018a). Learning to generalize: Meta-learning for domain generalization. In *AAAI*

Li, Y., Tian, X., Gong, M., Liu, Y., Liu, T., Zhang, K., & Tao, D. (2018b). Deep domain generalization via conditional invariant adversarial networks. In *ECCV*

Lin, C., Yuan, Z., Zhao, S., Sun, P., Wang, C., & Cai, J. (2021). Domain-invariant disentangled network for generalizable object detection. In *ICCV*

Liu, W., Rabinovich, A., & Berg, A. C. (2015). ParseNet: Looking wider to see better. arXiv preprint arXiv:1506.04579

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *ICLR*

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*

Neuhold, G., Ollmann, T., Rota Bulo, S., & Kontschieder, P. (2017). The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*

Nuriel, O., Benaim, S., & Wolf, L. (2021). Permuted AdaIN: Reducing the bias towards global statistics in image classification. In *CVPR*

Pan, X., Luo, P., Shi, J., & Tang, X. (2018). Two at once: Enhancing learning and generalization capacities via IBN-Net. In *ECCV*

Pan, X., Zhan, X., Shi, J., Tang, X., & Luo, P. (2019). Switchable whitening for deep representation learning. In *ICCV*

Peng, D., Lei, Y., Liu, L., Zhang, P., & Liu, J. (2021). Global and local texture randomization for synthetic-to-real semantic segmentation. *IEEE TIP, 30*, 6594–6608.

Peng, D., Lei, Y., Hayat, M., Guo, Y., & Li, W. (2022). Semantic-aware domain generalized segmentation. In *CVPR*

Qi, C. R., Yi, L., Su, H., & Guibas, L. J. (2017). PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*

Qiao, F., Zhao, L., & Peng, X. (2020). Learning to learn single domain generalization. In *CVPR*

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*

Richter, S. R., Vineet, V., Roth, S., & Koltun, V. (2016). Playing for data: Ground truth from computer games. In *ECCV*

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., & Lopez, A. M. (2016). The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*

Roy, S., Liu, M., Zhong, Z., Sebe, N., & Ricci, E. (2022). Class-incremental novel class discovery. In *ECCV*

Sakaridis, C., Dai, D., & Van Gool, L. (2018). Semantic foggy scene understanding with synthetic data. *IJCV, 126*, 973–992.

Sakaridis, C., Dai, D., & Gool, L. V. (2019). Guided curriculum model adaptation and uncertainty-aware evaluation for semantic night-time image segmentation. In *ICCV*

Sakaridis, C., Dai, D., & Van Gool, L. (2021). ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *ICCV*

Shankar, S., Piratla, V., Chakrabarti, S., Chaudhuri, S., Jyothi, P., & Sarawagi, S. (2018). Generalizing across domains via cross-gradient training. In *ICLR*

Shui, C., Li, Z., Li, J., Gagné, C., Ling, C. X., & Wang, B. (2021). Aggregating from multiple target-shifted sources. In *ICML*

Shui, C., Chen, Q., Wen, J., Zhou, F., Gagné, C., & Wang, B. (2022). A novel domain adaptation theory with Jensen–Shannon divergence. *Knowledge-Based Systems, 257*, 109808.

Shui, C., Wang, B., & Gagné, C. (2022). On the benefits of representation regularization in invariance based domain generalization. *Machine Learning, 111*, 895–915.

Shui, C., Xu, G., Chen, Q., Li, J., Ling, C. X., Arbel, T., Wang, B., & Gagné, C. (2022c). On learning fairness and accuracy on multiple subgroups. In *NeurIPS*

Tang, Z., Gao, Y., Zhu, Y., Zhang, Z., Li, M., & Metaxas, D. (2021). SelfNorm and CrossNorm for out-of-distribution robustness. In *ICCV*

Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Wang, H., Xiao, C., Kossaifi, J., Yu, Z., Anandkumar, A., & Wang, Z. (2021a). AugMax: Adversarial composition of random augmentations for robust training. In *NeurIPS*

Wang, P., Li, Y., & Vasconcelos, N. (2021b). Rethinking and improving the robustness of image style transfer. In *CVPR*

Wang, Z., Luo, Y., Qiu, R., Huang, Z., & Baktashmotlagh, M. (2021c). Learning to diversify for single domain generalization. In *ICCV*

Wu, A., & Deng, C. (2022). Single-domain generalized object detection in urban scene via cyclic-disentangled self-distillation. In *CVPR*

Wu, A., Liu, R., Han, Y., Zhu, L., & Yang, Y. (2021). Vector-decomposed disentanglement for domain-invariant object detection. In *ICCV*

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS, 34*, 12077–12090.

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., & Darrell, T. (2020). BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*

Yuan, J., Ma, X., Chen, D., Kuang, K., Wu, F., & Lin, L. (2022). Domain-specific bias filtering for single labeled domain generalization. *IJCV, 131*, 552–571.

Yue, X., Zhang, Y., Zhao, S., Sangiovanni-Vincentelli, A., Keutzer, K., & Gong, B. (2019). Domain randomization and pyramid consistency: simulation-to-real generalization without accessing target domain data. In *ICCV*

Zhao, L., Liu, T., Peng, X., & Metaxas, D. (2020). Maximum-entropy adversarial data augmentation for improved generalization and robustness. In *NeurIPS*

Zhao, Y., Zhong, Z., Yang, F., Luo, Z., Lin, Y., Li, S., & Nicu, S. (2021). Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification. In *CVPR*

Zhao, Y., Zhong, Z., Luo, Z., Lee, G. H., & Sebe, N. (2022). Source-free open compound domain adaptation in semantic segmentation. *IEEE TCSVT, 32*, 7019–7032.

Zhao, Y., Zhong, Z., Sebe, N., & Lee, G. H. (2022b). Novel class discovery in semantic segmentation. In *CVPR*

Zhao, Y., Zhong, Z., Zhao, N., Sebe, N., & Lee, G. H. (2022c). Style-hallucinated dual consistency learning for domain generalized semantic segmentation. In *ECCV*

Zheng, Z., & Yang, Y. (2020). Unsupervised scene adaptation with memory regularization in vivo. In *IJCAI*

Zheng, Z., & Yang, Y. (2021). Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. In *IJCV*

Zheng, Z., & Yang, Y. (2022). Adaptive boosting for domain adaptation: Toward robust predictions in scene segmentation. *IEEE TIP, 31*, 5371–5382.

Zhong, Z., Zhu, L., Luo, Z., Li, S., Yang, Y., & Sebe, N. (2021). Open-Mix: Reviving known knowledge for discovering novel visual categories in an open world. In *CVPR*

Zhong, Z., Zhao, Y., Lee, G. H., & Sebe, N. (2022). Adversarial style augmentation for domain generalized urban-scene segmentation. In *NeurIPS*

Zhou, K., Yang, Y., Qiao, Y., & Xiang, T. (2021a). Domain generalization with mixstyle. In *ICLR*

Zhou, Q., Feng, Z., Gu, Q., Pang, J., Cheng, G., Lu, X., Shi, J., & Ma, L. (2021b). Context-aware mixup for domain adaptive semantic segmentation. arXiv preprint arXiv:2108.03557

Zhou, Q., Feng, Z., Gu, Q., Cheng, G., Lu, X., Shi, J., & Ma, L. (2022). Uncertainty-aware consistency regularization for cross-domain semantic segmentation. *Computer Vision and Image Understanding, 221*, 103448.

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*