International Journal of Computer Vision
https://doi.org/10.1007/s11263-023-01878-8

Intra- & Extra-Source Exemplar-Based Style Synthesis for Improved Domain Generalization

Yumeng Li[1,2] · **Dan Zhang**[1,3] · **Margret Keuper**[2,4] · **Anna Khoreva**[1,3]

Received: 29 November 2022 / Accepted: 10 August 2023

Abstract

The generalization with respect to domain shifts, as they frequently appear in applications such as autonomous driving, is one of the remaining big challenges for deep learning models. Therefore, we propose an exemplar-based style synthesis pipeline to improve domain generalization in semantic segmentation. Our method is based on a novel masked noise encoder for StyleGAN2 inversion. The model learns to faithfully reconstruct the image, preserving its semantic layout through noise prediction. Random masking of the estimated noise enables the style mixing capability of our model, i.e. it allows to alter the global appearance without affecting the semantic layout of an image. Using the proposed masked noise encoder to randomize style and content combinations in the training set, i.e., intra-source style augmentation (ISSA) effectively increases the diversity of training data and reduces spurious correlation. As a result, we achieve up to 12.4% mIoU improvements on driving-scene semantic segmentation under different types of data shifts, i.e., changing geographic locations, adverse weather conditions, and day to night. ISSA is model-agnostic and straightforwardly applicable with CNNs and Transformers. It is also complementary to other domain generalization techniques, e.g., it improves the recent state-of-the-art solution RobustNet by 3% mIoU in Cityscapes to Dark Zürich. In addition, we demonstrate the strong plug-n-play ability of the proposed style synthesis pipeline, which is readily usable for extra-source exemplars e.g., web-crawled images, without any retraining or fine-tuning. Moreover, we study a new use case to indicate neural network's generalization capability by building a stylized proxy validation set. This application has significant practical sense for selecting models to be deployed in the open-world environment. Our code is available at https://github.com/boschresearch/ISSA.

Keywords Domain generalization · GAN inversion · Data augmentation · Semantic segmentation

1 Introduction

The varying environment with potentially diverse illumination and adverse weather conditions makes challenging the deployment of deep learning models in an open-world (Sakaridis et al., 2021; Zhang et al., 2021a). Therefore, improving the generalization capability of neural networks is crucial for safety-critical applications such as autonomous driving (see for example Fig. 1. While generally the target domains can be inaccessible or unpredictable at training time, it is important to train a generalizable model, based on the known (source) domain, which may offer only a limited or biased view of the real world (Burton et al., 2017; Shafaei et al., 2018).

Diversity of the training data is considered to play an important role for domain generalization, including natural distribution shifts (Taori et al., 2020). Many existing works assume that multiple source domains are accessible during

Communicated by Bumsub Ham.

✉ Yumeng Li
yumeng.li@bosch.com

Dan Zhang
dan.zhang2@bosch.com

Margret Keuper
margret.keuper@uni-siegen.de

Anna Khoreva
anna.khoreva@bosch.com

[1] Bosch Center for Artificial Intelligence, Renningen, Germany

[2] University of Siegen, Siegen, Germany

[3] University of Tübingen, Tübingen, Germany

[4] Max Planck Institute for Informatics, Saarbrücken, Germany

Published online: 07 September 2023

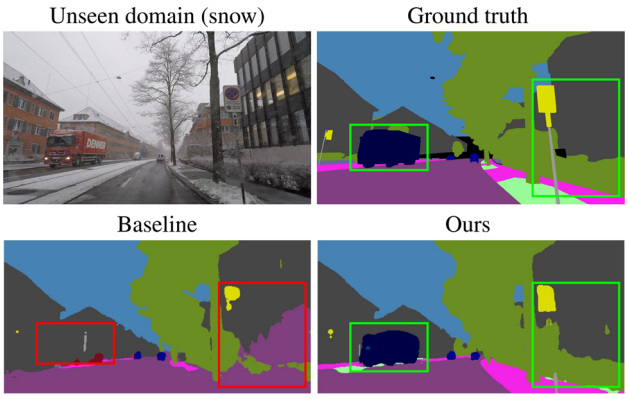

Fig. 1 Semantic segmentation results of HRNet (Wang et al., 2021b) on unseen domain (snow), trained on Cityscapes (Cordts et al., 2016) and tested on ACDC (Sakaridis et al., 2021). The model trained with our ISSA can successfully segment the truck, while the baseline model fails completely

training (Hu et al., 2020; Li et al., 2018a; Balaji et al., 2018; Li et al., 2018b, 2020; Jin et al., 2020; Zhou et al., 2020). For instance, Li et al. (2018a) applied meta-learning to better generalize to unseen domains, where source domains are divided into meta-source and meta-target domains to simulate domain shift; Hu et al. (2020) propose multi-domain discriminant analysis to learn a domain-invariant feature transformation. However, for pixel-level prediction tasks such as semantic segmentation, collecting diverse training data involves a tedious and costly annotation process (Caesar et al., 2018). Therefore, improving and predicting generalization from a *single source domain* is exceptionally compelling, particularly for semantic segmentation.

One pragmatic way to improve data diversity is by applying data augmentation. It has been widely adopted in solving different tasks, such as image classification (Zhang et al., 2018a; Zhou et al., 2021; Hendrycks et al., 2019; Verma et al., 2019; Hong et al., 2021), GAN training with limited data (Karras et al., 2020a; Jiang et al., 2021), or pose estimation (Peng et al., 2018; Bin et al., 2020; Wang et al., 2021a). One line of data augmentation techniques focuses on increasing the content diversity in the training set, such as geometric transformation (e.g., cropping or flipping), CutOut (DeVries & Taylor, 2017), and CutMix (Yun et al., 2019). However, CutOut and CutMix are ineffective on natural domain shifts, as reported in (Taori et al., 2020). Style augmentation, on the other hand, only modifies the style—the non-semantic appearance such as texture and color of the image (Gatys et al., 2016)—while preserving the semantic content. By diversifying the style and content combinations, style augmentation can reduce overfitting to the style-content correlation in the training set, improving robustness against domain shifts. Hendrycks corruptions (Hendrycks & Dietterich, 2019) provide a wide range of synthetic styles, including weather conditions. However, they are not always realistic looking, thus being still far from resembling natural data shifts. In this work, we propose an exemplar-based style synthesis pipeline for semantic segmentation, aiming to improve the style diversity in the training and validation set without extra labeling effort.

Our exemplar-based style synthesis technique is based on the inversion of StyleGAN2 (Karras et al., 2020b), which is the state-of-the-art unconditional Generative Adversarial Network (GAN) and thus ensures high quality and realism of synthetic samples. GAN inversion allows encoding a given image to latent variables, and thus facilitates faithful reconstruction with style mixing capability. To realize the synthesis pipeline, we learn to separate semantic content from style information based on a single source domain. This allows to alter the style of an image while leaving the content unchanged. In particular, we focus on intra-source style augmentation (ISSA). Namely, our exemplar-based style synthesis makes use of training samples from the source domain, extracting their styles and contents followed by randomly mixing them up. In doing so, we can increase the data diversity and alleviate the spurious correlation in the given training data.

The faithful reconstruction of images with complex structures such as driving scenes is non-trivial. Prior methods (Richardson et al., 2021; Yao et al., 2022; Roich et al., 2022; Alaluf et al., 2022; Dinh et al., 2022; Šubrtová et al., 2022) are mainly tested on simple single-object-centric datasets, e.g., FFHQ (Karras et al., 2019), CelebA-HQ (Karras et al., 2018), or LSUN (Yu et al., 2015). As shown in (Abdal et al., 2020), extending the native latent space of StyleGAN2 with a stochastic noise space can lead to improved inversion quality. However, all style *and* content information will be embedded in the noise map, leaving the latent codes inactive in this setting. Therefore, to enable the precise reconstruction of complex driving scenes as well as style mixing, we propose a masked noise encoder for StyleGAN2. The proposed random masking regularization on the noise map encourages the generator to rely on the latent prediction for reconstruction. Thus, it allows to effectively separate content and style information and facilitates realistic style mixing, as shown in Fig. 2.

We further discover an excellent plug-n-play ability of the proposed style synthesis pipeline, i.e., it can be directly applied to unseen domains without requiring the re-training of the encoder or generator. For instance, in Fig. 11, we employ our pipeline directly on web-crawled images, where the model is only trained on Cityscapes. This appealing property opens up the opportunity to go beyond intra-source exemplar-based style mixing, and grants us more flexibility to harness extra-source data for style synthesis. Thus, we also experiment with extra-source style argumentation (ESSA) to further improve the generalization performance.

Besides data augmentation, we explore the usage of the proposed pipeline for assessing neural networks' generalization capability in Sect. 6. By transferring styles from unannotated data samples of the target domain to existing labelled data, we can build a style-augmented proxy set for validation without introducing extra-labelling effort. We observe that performance on this proxy set has a strong correlation with the real test performance on unseen target data, which could be used in practice to select more suitable models for deployment.

In summary, we make the following contributions:

- We propose a masked noise encoder for GAN inversion, which enables high quality reconstruction and style mixing of complex scene-centric datasets.
- We exploit GAN inversion for intra-source data augmentation, which can improve generalization under natural distribution shifts on semantic segmentation.
- Extensive experiments demonstrate that our proposed augmentation method ISSA consistently promotes domain generalization performance on driving-scene semantic segmentation across different network architectures, achieving up to 12.4% mIoU improvement, even with limited diversity in the source data and without access to the target domain.
- We discover the plug-n-play ability of our masked noise encoder, and showcase its potential of direct application on extra-source data such as web-crawled images.
- We further explore the usage of the proposed pipeline for assessing models' generalization performance on unseen data. By building a style-augmented proxy validation set on known labelled data, we observe that there is a strong correlation between the performance on the proxy validation set and the real test set, which offers useful insights for model selection without introducing any extra annotation effort.

This paper is an extended version of our previous work (Li et al., 2023) with more experimental evaluation and discussion on the potential and two new applications of the proposed method. In particular, we provide a more detailed ablation study on the design of the proposed masked noise encoder (see Tables 3 and 4, Fig. 8. Furthermore, we add a discussion on the plug-n-play ability of the pipeline and go beyond intra-source domain to extra-source domain style mixing. We also conducted new experiments reported in Tables 11 and 12. Finally, the new application as model generalization performance indicator is introduced in Sect. 6.

2 Related Work

Domain Generalization Domain generalization concerns the generalization ability of neural networks to a target domain that follows a different distribution than the source domain, and prior knowledge of the target domain is inaccessible at training. Various methods have been proposed to approach this problem from different angles, which employ data augmentation (Khirodkar et al., 2019; Somavarapu et al., 2020; Huang et al., 2021; Zhou et al., 2021; Li et al., 2022), domain alignment (Hu et al., 2020; Li et al., 2020; Jin et al., 2020; Zhou et al., 2020), adversarial training (Li et al., 2018b; Shao et al., 2019; Rahman et al., 2020; Deng et al., 2020), meta-learning (Li et al., 2018a; Balaji et al., 2018; Li et al., 2019a; Zhao et al., 2021), ensemble learning (D'Innocente & Caputo, 2018; Mancini et al., 2018; Wu & Gong, 2021; Lee et al., 2022a), or feature decomposition (Wan et al., 2022; Chen et al., 2022). Particularly, (Qiao et al., 2020; Wang et al., 2021c; Jia et al., 2020; Ouyang et al., 2022) focus on single domain generalization problem. While the majority focuses on image-level tasks, e.g., image classification or person re-identification, a few recent works (Choi et al., 2021; Lee et al., 2022b; Kim et al., 2023, 2022; Zhao et al., 2022) investigate pixel-level prediction tasks such as semantic segmentation. RobustNet (Choi et al., 2021) proposes an instance selective whitening loss to the instance normalization, aiming to selectively remove information that causes a domain shift while maintaining discriminative features. (Kim et al., 2022) introduces a memory-guided meta-learning framework to capture co-occurring categorical knowledge across domains. (Lee et al., 2022b; Kim et al., 2023) make use of extra data in the wild for feature augmentation. SHADE (Zhao et al., 2022) proposed to use a style consistency constraint to learn a style-invariant representation and a retrospection consistency constraint to leverage knowledge from the pretrained backbone. To assist the training, they perturb features to simulate style variations.

Another line of work explores feature-level augmentation (Zhou et al., 2021; Li et al., 2022). MixStyle (Zhou et al., 2021) and DSU (Li et al., 2022) add perturbation at the normalization layer to simulate domain shifts at test time. However, this perturbation can potentially cause a distortion of the image content, which can be harmful for semantic segmentation (see Sect. 4.3. Moreover, these methods require a careful adaptation to the specific network architecture. In contrast, ISSA performs style mixing on the image-level, thus being model-agnostic, and can be applied as a complement to other methods in order to further increase the generalization performance.

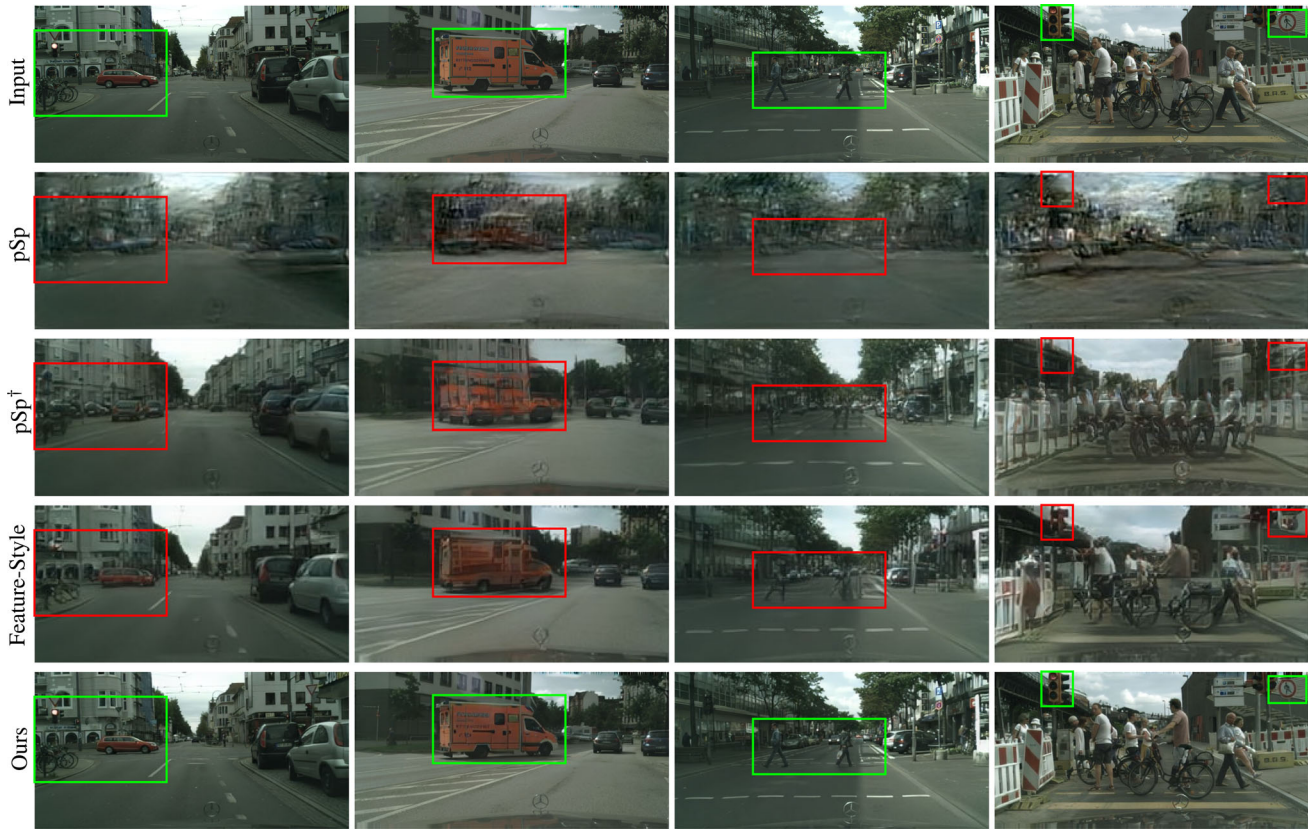

Fig. 2 Qualitative results (best view in color and zoom in) of StyleGAN2 inversion methods on Cityscapes, i.e., pSp (Richardson et al., 2021), pSp†, Feature-Style encoder (Yao et al., 2022) and our masked noise encoder. Note, pSp† is an improved version of pSp (Richardson et al., 2021) introduced by us, training pSp with an additional discriminator and incorporate synthesized images for better initialization. pSp† can reconstruct the rough layout of the scene but still struggles to preserve details. The Feature-Style encoder shows a better reconstruction quality, yet it cannot faithfully reconstruct small objects (e.g. pedestrian), and some objects (e.g. the vehicle, bicycle) are rather blurry. Our masked noise encoder has highest image fidelity, preserving finer details in the inverted image (Color figure online)

Beyond data augmentation for improving domain generalization, we further explore the usage of our exemplar-based style synthesis pipeline for assessing the generalization performance. Recently, Zhang et al. (2021b) proposed to predict generalization of image classifiers using performance on synthetic data produced by a conditional GAN. While this is limited to the generalization in the source domain, and it is not straightforward how to apply it on semantic segmentation task. In contrast to generating image from scratch, We employ proposed exemplar-based style synthesis pipeline to augment labelled source data and build a stylized proxy validation sets. We empirically show that such proxy validation sets can indicate generalization performance, without extra annotation required.

Data Augmentation Data augmentation techniques can diversify training samples by altering their style, content, or both, thus preventing overfitting and improving generalization. Mixup augmentations (Zhang et al., 2018a; Dabouei et al., 2021; Verma et al., 2019) linearly interpolate between two training samples and their labels, regularizing both style and content. Despite effectiveness shown on image-level classification tasks, they are not well suited for dense pixel-level prediction tasks. CutMix (Yun et al., 2019) cuts and pastes a random rectangular region of the input image into another image, thus increasing the content diversity. Geometric transformation, e.g., random scaling and horizontal flipping, can also serve this purpose. In contrast, Hendrycks corruptions (Hendrycks & Dietterich, 2019) only affect the image appearance without modifying the content. Their generated images look artificial, being far from resembling natural data, and thus offer limited help against natural distribution shifts (Taori et al., 2020).

StyleMix (Hong et al., 2021) is conceptually closer to our method, which aims to decompose training images into content and style representations and then mix them up to generate more samples. Nonetheless, their AdaIN (Huang & Belongie, 2017) based style mixing method cannot fulfill the pixel-wise label-preserving requirement (see Fig. 10. Another line of CycleGAN based style transfer methods

(Hoffman et al., 2018; Voreiter et al., 2020) require the access to both source and target domain during training, and thus cannot be employed for domain generalization problem, where the target domains remain unknown during training time. Our ISSA is also a style-based data augmentation technique that leverages the capabilities of a state-of-the-art GAN to produce natural looking samples. By modifying solely the style of the input images and maintaining their content intact, the original ground truth label maps can be reused. Importantly, this model can be effectively trained on a single domain without necessitating target data. This is a crucial property, when employed for data augmentation and to enhance other network's generalization performance.

GAN Inversion Showing good results, GAN inversion has been explored for many applications such as face editing (Abdal et al., 2019, 2020; Zhu et al., 2020), image restoration (Pan et al., 2022), and data augmentation (Nguyen et al., 2021; Golhar et al., 2022). StyleGANs (Karras et al., 2019, 2020a, b) are commonly used for inversion, as they demonstrate high synthesis quality and appealing editing capabilities. Nevertheless, there is a known distortion-editability trade-off (Tov et al., 2021). Thus, it is crucial to achieve a curated performance for a specific use case.

GAN inversion approaches can be classified into three groups: optimization based methods (Creswell & Bharath, 2019; Abdal et al., 2019, 2020; Gu et al., 2020; Kang et al., 2021; Collins et al., 2020), encoder based models (Richardson et al., 2021; Yao et al., 2022; Bartz et al., 2021; Tov et al., 2021; Wei et al., 2022) methods, and hybrid approaches (Dinh et al., 2022; Roich et al., 2022; Alaluf et al., 2022; Chai et al., 2021; Song et al., 2022). Optimization methods generally have worse editability and need exhaustive optimization for each input. Thus, in this paper, we use an encoder based method for our style mixing purpose. The representative encoder based work pSp encoder (Richardson et al., 2021) embeds the input image in the extended latent space $\mathcal{W}^+$ of StyleGAN. The e4e encoder (Tov et al., 2021) improves editability of pSp while trading off detail preservation. Yet, for the semantic segmentation augmentation task, it is crucial to assure the pixel-wise alignment with ground-truth label maps. To improve the reconstruction quality, the Feature-Style encoder (Yao et al., 2022) further replaces the lower latent code prediction with a feature map prediction. Recent works explored the usage of additional information such as labelled regions of interest (Moon & Park, 2022) and segment masks (Šubrtová et al., 2022), or involved the joint optimization of the generator (Roich et al., 2022; Hu, 2022). Our method only requires RGB images and a frozen generator, meanwhile offers plug-n-play ability on web-crawled images (see Sect. 5).

Despite much progress, most prior work only showcases applications on single object-centric datasets, such as CelebA-HQ (Karras et al., 2018), FFHQ (Karras et al., 2019), LSUN (Yu et al., 2015). They still fail on more complex scenes, thus restricting their application in practice. Our masked noise encoder can fulfil both the fidelity and the style mixing capability requirements, rendering itself well-suited for data augmentation for semantic segmentation. To the best of our knowledge, our approach is the first GAN inversion method which can be effectively applied as data augmentation for the semantic segmentation of complex scenes.

3 Method

We introduce our exemplar-based style synthesis pipeline in Sect. 3.1, which relies on GAN inversion that can offer faithful reconstruction and style mixing of images. To enable better style-content disentanglement, we propose a masked noise encoder for GAN inversion in Sect. 3.2. Its detailed training loss is described in Sect. 3.3.

3.1 Exemplar-Based Style Synthesis Pipeline

The lack of data diversity and the existence of spurious correlation in the training set often lead to poor domain generalization. To mitigate them, the proposed style synthesis pipeline aims at (1) extracting styles from given exemplars, and (2) augmenting the training samples in the source domain with the new styles, while preserving their semantic content. For data augmentation, it employs GAN inversion to randomize the style-content combinations. In doing so, it diversifies the source dataset and reduces spurious style-content correlations. Because the content of images is preserved and only the style is changed, the ground truth label maps can be reused for training and validation, without requiring any further annotation effort.

Our style synthesis pipeline is built on top of an encoder-based GAN inversion, given its fast inference. GANs, such as StyleGANs (Karras et al., 2019, 2020a, b), have shown the capability of encoding rich semantic and style information in intermediate features and latent spaces. For encoder-based GAN inversion, an encoder is trained to invert an input image back into the latent space of a pre-trained GAN generator. The encoder is desired to separately encode the style and content information of the input image. With such an encoder, it can synthesize new training samples with new style-content combinations. In particular, we are interested in intra-source style augmentation (ISSA), where the encoder should take the content and style codes from different training samples within the source domain and feed them to the pre-trained generator. If this encoder-based GAN inversion can also handle unseen data, we will further make use the styles of exemplars outside the source domain, such as web-crawled images, enabling extra-source style augmentation (ESSA). In both cases, since

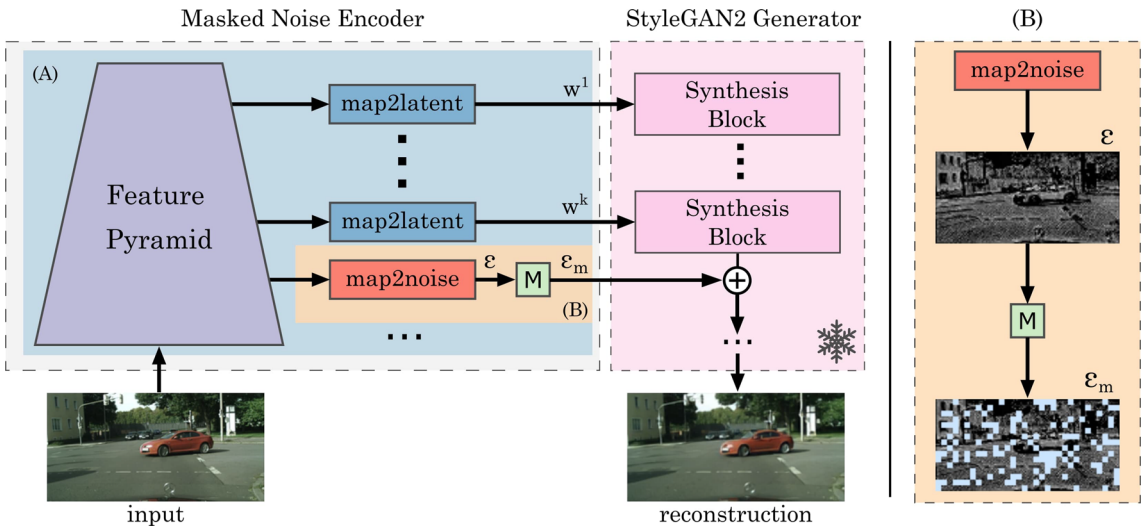

Fig. 3 Method overview. Our encoder is built on top of the pSp encoder (Richardson et al., 2021), shown in the blue area (**A**). It maps the input image to the extended latent space $\mathcal{W}^+$ of the pre-trained StyleGAN2 generator. To promote the reconstruction quality on complex scene-centric dataset, e.g., Cityscapes, our encoder additionally predicts the noise map at an intermediate scale, illustrated in the orange area (**B**). M stands for random noise masking, regularization for the encoder training. Without it, the noise map overtakes the latent codes in encoding the image style, so that the latter cannot make any perceivable changes on the reconstructed image, thus making style mixing impossible (Color figure online)

only the styles of the training samples in the source domain are modified, the newly synthesized training samples already have their ground truth label maps in place.

StyleGAN2 can synthesize natural looking images resembling scene-centric datasets such as Cityscapes (Cordts et al., 2016) and BDD100K (Yu et al., 2020). However, existing GAN inversion encoders cannot provide the desired fidelity and style mixing capability to enable ISSA and ESSA for an improved domain generalization of semantic segmentation. Loss of fine details or inauthentic reconstruction of small-scale objects would even harm the model's generalization ability. Therefore, we propose a novel encoder design to invert StyleGAN2, termed *masked noise encoder* (see Fig. 3).

3.2 Masked Noise Encoder

We build our encoder upon the pSp encoder (Richardson et al., 2021). It employs a feature pyramid (Lin et al., 2017) to extract multi-scale features from a given image, see Fig. 3A. We improve over pSp by identifying in which latent space to embed the input image for the high-quality reconstruction of the images with complex street scenes. Further, we propose a novel training scheme to enable the style-content disentanglement of the encoder, thus improving its style mixing capability.

Extended Latent Space The StyleGAN2 generator takes the latent code $w \in \mathcal{W}$ generated by an MLP network and randomly sampled additive Gaussian noise maps $\{\epsilon\}$ as inputs for image synthesis. As pointed out in (Abdal et al., 2019), it is suboptimal to embed a real image into the original latent space $\mathcal{W}$ of StyleGAN2, due to the gap between the real and synthetic data distributions. A common practice is to map the input image into the extended latent space $\mathcal{W}^+$. The multi-scale features of the pSp feature pyramid are respectively mapped to the latent codes $\{w^k\}$ at the corresponding scales of the StyleGAN2 generator, i.e., map2latent in Fig. 3A.

Additive Noise Map The latent codes $\{w^k\}$ from the extended latent space $\mathcal{W}^+$ alone are not expressive enough to reconstruct images with diverse semantic layouts such as Cityscapes (Cordts et al., 2016) as shown in Fig. 2-(pSp†). The latent codes of StyleGAN2 are one-dimensional vectors that modulate the feature vectors at different spatial positions identically. Therefore, they cannot precisely encode the semantic layout information, which is spatially varying. To address this issue, our encoder additionally predicts the additive noise map ε of the StyleGAN2 at an intermediate scale, i.e., map2noise in Fig. 3B. The noise map ε has spatial dimensions, making it inherently capable of encoding more information. It is particularly advantageous when dealing with content information that varies spatially, as the noise map can more readily accommodate such information. As evidenced by the visualization presented in Fig. 5, the noise map is adept at capturing the semantic content of the scene.

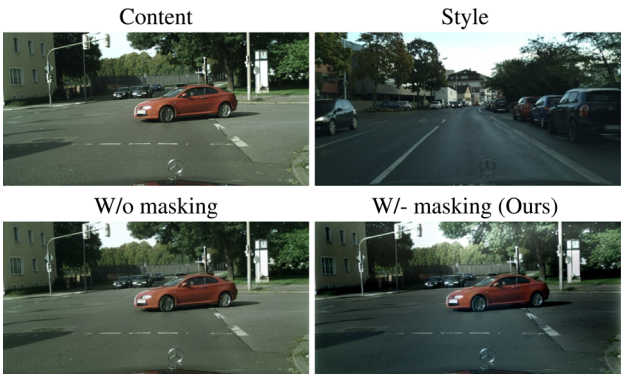

Fig. 4 Style mixing effect enabled by random noise masking (best view in color). Despite the good reconstruction quality, the encoder trained without masking cannot change the style of the given Content image. In contrast, the encoder trained with masking can modify it using the style from the given Style image (Color figure online)

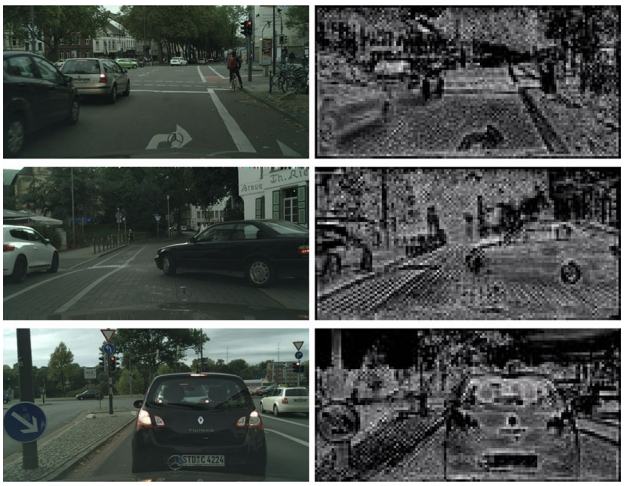

Fig. 5 Noise map visualization of our masked noise encoder. The noise map encodes the semantic content of the image

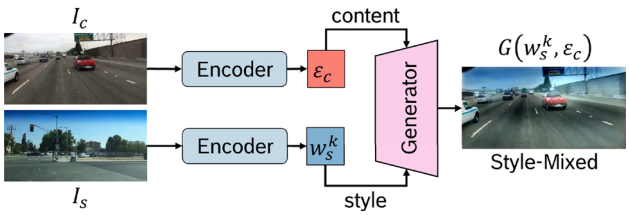

Fig. 6 Style mixing process. The generator G takes the latent codes $\{w_s^k\}$ of I_s and the noise map ε_c of I_c, and produce the stylized image, i.e., $G(w_s^k, \varepsilon_c)$.

Random Noise Masking While offering high-quality reconstruction, the additive noise map can be too expressive so that it encodes nearly all perceivable details of the input image. This results in a poor style-content disentanglement and can damage the style mixing capability of the encoder (see Fig. 4). To avoid this undesired effect, we propose to regularize the noise prediction of the encoder by random masking of the noise map. Note that the random masking as a regularization technique has also been successfully used in reconstruction-based self-supervised learning (Xie et al., 2022; He et al., 2022). In particular, we spatially divide the noise map into non-overlapping $P \times P$ patches, see M in Fig. 3B. Based on a pre-defined ratio ρ, a subset of patches is randomly selected and replaced by patches of unit Gaussian random variables $\epsilon \sim N(0, 1)$ of the same size. $N(0, 1)$ is the prior distribution of the noise map at training the StyleGAN2 generator. We call this encoder *masked noise encoder* as it is trained with random masking to predict the noise map.

The proposed random masking reduces the encoding capacity of the noise map, hence encouraging the encoder to jointly exploit the latent codes $\{w^k\}$ for reconstruction. Figure 7 visualizes the style mixing effect. The encoder takes the noise map ε_c and latent codes $\{w_s^k\}$ from the content image and style image, respectively. Then, they are fed into StyleGAN2 to synthesize a new image, i.e., $G(w_s^k, \varepsilon_c)$, as illustrated in Fig. 6. If the encoder is not trained with random masking, the new image does not have any perceptible difference with the content image. This means the latent codes $\{w^k\}$ encode negligible information of the image. In contrast, when being trained with masking, the encoder creates a novel image that takes the content and style from two different images. This observation confirms the enabling role of masking for content and style disentanglement, and thus the improved style mixing capability. The noise map no longer encodes all perceptible information of the image, including style and content. In effect, the latent codes $\{w^k\}$ play a more active role in controlling the style. In Fig. 5, we further visualize the noise map of the masked noise encoder and observe that it captures well the semantic content of the scene.

Additionally, we discover that our masked noise encoder is equipped with strong plug-n-play ability, i.e., readily usable on novel domains without retraining or fine-tuning. As shown in Fig. 11, the masked noise encoder together with the generator which is trained on Cityscapes not only reconstruct unseen domain data (e.g., north polar bear), but also remain the style mixing capability (e.g., turning bright day into a sunset scene). This generalization capability allows us to further exploit extra-source data for style synthesis, i.e., ESSA. Except that the styles are extracted from external exemplars, the style synthesis process of ESSA is identical to ISSA.

3.3 Encoder Training Loss

Mathematically, the proposed StyleGAN2 inversion with the masked noised encoder E^M can be formulated as

$$\{w^1, \ldots, w^K, \varepsilon\} = E^M(x); \qquad (3.1)$$
$$x^* = G \circ E^M(x) = G(w^1, \ldots, w^K, \varepsilon).$$

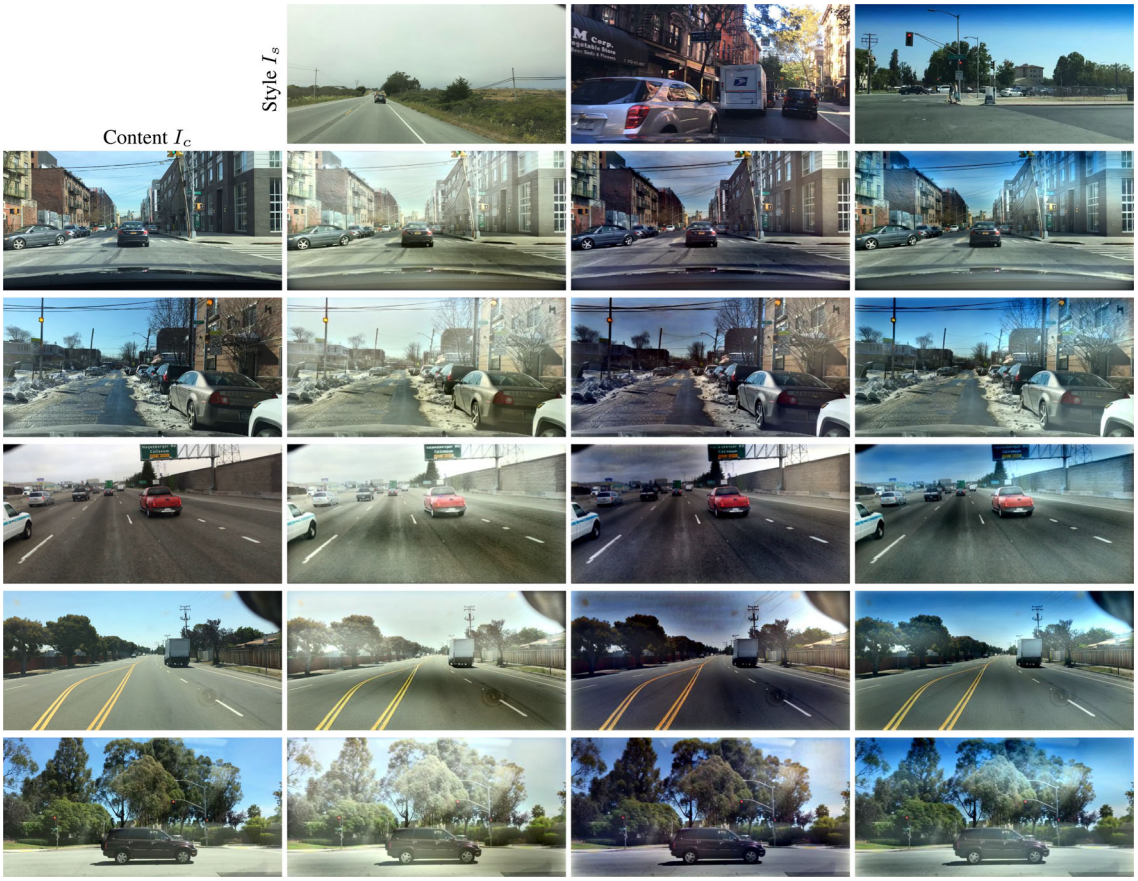

Fig. 7 Visual examples of style mixing on BDD100K (best view in color) enabled by our masked noise encoder. By combining the latent codes $\{w_s^k\}$ of I_s and the noise map ε_c of I_c, the synthesized images $G(w_s^k, \varepsilon_c)$ preserve the content of I_c with a new style resembling I_s (Color figure online)

The masked noise encoder E^M maps the given image x onto the latent codes $\{w^k\}$ and the noise map ε. The StyleGAN2 generator G takes both $\{w^k\}$ and ε as the input and generates x^*. Ideally, x^* should be identical to x, i.e., a perfect reconstruction.

When training the masked noise encoder E^M to reconstruct x, the original noise map ε is masked before being fed into the pre-trained G

$$\varepsilon_M = (1 - M_{noise}) \odot \varepsilon + M_{noise} \odot \epsilon, \tag{3.2}$$

$$\tilde{x} = G(w^1, \ldots, w^K, \varepsilon_M), \tag{3.3}$$

where M_{noise} is the random binary mask, $\odot$ indicates the Hadamard product, and $\epsilon \sim N(0, 1)$ is the random Gaussian noise. $\tilde{x}$ denotes the reconstructed image with the masked noise ε_M. The training loss for the encoder is given as

$$\mathcal{L} = \mathcal{L}_{mse} + \lambda_1 \mathcal{L}_{lpips} + \lambda_2 \mathcal{L}_{adv} + \lambda_3 \mathcal{L}_{reg}, \tag{3.4}$$

where $\{\lambda_i\}$ are weighting factors. The first three terms are the pixel-wise MSE loss, learned perceptual image patch similarity (LPIPS) (Zhang et al., 2018b) loss and adversarial loss (Goodfellow et al., 2014),

$$\mathcal{L}_{mse} = \left\| (1 - M_{img}) \odot (x - \tilde{x}) \right\|_2, \tag{3.5}$$

$$\mathcal{L}_{lpips} = \left\| (1 - M_{feat}) \odot (\mathrm{VGG}(x) - \mathrm{VGG}(\tilde{x})) \right\|_2, \tag{3.6}$$

$$\mathcal{L}_{adv} = -\log D(G(E^M(x))). \tag{3.7}$$

which are the common reconstruction losses for encoder training (Richardson et al., 2021; Zhu et al., 2020). Note that masking removes the information of the given image x at certain spatial positions, the reconstruction requirement on these positions should then be relaxed. M_{img} and M_{feat} are obtained by up- and down-sampling the noise mask M_{noise} to the image size and the feature size of the VGG-based feature extractor. The adversarial loss is obtained by formulating the encoder training as an adversarial game with a discriminator D that is trained to distinguish between reconstructed and real images.

The last regularization term is defined as

$$\mathcal{L}_{reg} = \|\varepsilon\|_1 + \left\| E_w^M(G(w_{gt}, \epsilon)) - w_{gt} \right\|_2. \tag{3.8}$$

The L1 norm helps to induce sparse noise prediction. It is complementary to random masking, reducing the capacity of the noise map. The second term is obtained by using the ground truth latent codes w_{gt} of synthesized images $G(w_{gt}, \epsilon)$ to train the latent code prediction $E_w^M(\cdot)$ (Yao et al., 2022). It guides the encoder to stay close to the original latent space of the generator, speeding up the convergence.

4 Experiments

We start from the experiment setup in Sect. 4.1. Then, Sects. 4.2 and 4.3 respectively report our experiments on the masked noise encoder for StyleGAN2 inversion and ISSA for improved domain generalization of semantic segmentation.

4.1 Experiment Setup

Datasets We conduct extensive experiments on four driving scene datasets, which are Cityscapes (CS) (Cordts et al., 2016), BDD100K (BDD) (Yu et al., 2020), ACDC (Sakaridis et al., 2021) and Dark Zürich (DarkZ) (Sakaridis et al., 2019). Cityscapes is collected from different cities primarily in Germany, under good/medium weather conditions during daytime. BDD100K is a driving-scene dataset collected in the US, representing a geographic location shift from Cityscapes. Besides, it also includes more diverse scenes (e.g., city streets, residential areas, and highways) and different weather conditions captured at different times of the day. Both ACDC and Dark Zürich are collected in Switzerland. ACDC contains four adverse weather conditions (rain, fog, snow, night) and Dark Zürich contains night scenes. The default setting is to use Cityscapes as the source training data, whereas the validation sets of the other datasets represent unseen target domains with different types of natural shifts, i.e., used only for testing. Additionally, we also study the challenging day-to-night generalization scenario, where BDD100K-Daytime is used as the source set, ACDC-Night and Dark Zürich are treated as unseen domains. In both cases, we consider a *single source domain* for training.

Training Details We experiment with two image resolutions: 128×256 and 256×512. The StyleGAN2 (Karras et al., 2020a) model is first trained to *unconditionally* synthesize images and then fixed during the encoder training. To invert the pre-trained StyleGAN2 generator, the masked noise encoder predicts both latent codes in the extended $\mathcal{W}^+$ space and the additive noise map. In accordance with the StyleGAN2 generator, $\mathcal{W}^+$ space consists of 14 and 16 latent code vectors for the input resolution 128×256 and 256×512, respectively. The additive noise map is always at the intermediate feature space with one fourth of the input resolution. We use the same encoder architecture, optimizer, and learning rate scheduling as pSp (Richardson et al., 2021). Our encoder is trained with the loss function defined in Eq. (3.4) with $\lambda_1 = 10$ and $\lambda_2 = \lambda_3 = 0.1$. For our random noise masking, we use a patch size P of 4 with a masking ratio $\rho = 25\%$. A detailed ablation study on the masking and noise map of the encoder can be found in Sect. 4.2.

We use the trained masked noise encoder to perform ISSA as described in Sect. 3.1. We experiment with several architectures for semantic segmentation, i.e., HRNet (Wang et al., 2021b), SegFormer (Xie et al., 2021), and DeepLab v2/v3+ (Chen et al., 2017a, 2018). The baseline segmentation models are trained with their default configurations and using the standard augmentation, i.e., random scaling and horizontal flipping.

4.2 Masked Noise Encoder

Reconstruction quality Table 1 shows that our masked noise encoder considerably outperforms two strong StyleGAN2 inversion baselines pSp (Richardson et al., 2021) and Feature-Style encoder (Yao et al., 2022) in all three evaluation metrics. The achieved low values of MSE, LPIPS (Zhang et al., 2018b) and FID (Heusel et al., 2017) indicate its high-quality reconstruction. Both the masked noise encoder and the Feature-Style encoder adopt the adversarial loss $\mathcal{L}_{adv}$ and regularization using synthesized images with ground truth latent codes w_{gt}. Therefore, we also add them to train pSp and note this version as pSp†. While pSp† improves over pSp in MSE and FID, it still underperforms compared to the others. This confirms that inverting into the extended latent space $\mathcal{W}^+$ only allows limited reconstruction quality on Cityscapes. The Feature-Style encoder (Yao et al., 2022) replaces the prediction of the low level latent codes with feature prediction, which results in better reconstruction without severely harming style editability. However, its reconstruction on Cityscapes is still not satisfying and underperforms to our masked noise encoder. As noted in (Yao et al., 2022), the feature size of the Feature-Style encoder is restricted. Using a larger feature map to improve reconstruction quality can only be done as a replacement of more latent code predictions. Consequently, it largely reduces the expressiveness of the latent embedding and leads to extremely poor editability, being no longer suitable for downstream applications, e.g., style mixing data augmentation.

The visual comparison across pSp†, the Feature-Style encoder and our masked noise encoder is shown in Fig. 2 and is aligned with the quantitative results in Table 1. pSp† has overall poor reconstruction quality. The Feature-Style encoder cannot faithfully reconstruct small objects and restore fine details. In comparison, our masked noise encoder offers high-quality reconstruction, preserving the semantic layout and fine details of each class. Having a high-quality reconstruction is an important requirement for using the

Table 1 Reconstruction quality on Cityscapes at the resolution 128 × 256

Method	MSE ↓	LPIPS ↓	FID ↓
pSp (Richardson et al., 2021)	0.078	0.348	130.62
pSp† (Richardson et al., 2021)	0.049	0.339	14.60
Feature-Style (Yao et al., 2022)	0.025	0.220	7.14
Ours	**0.011**	**0.124**	**3.94**

Bold values indicate the best performance
MSE, LPIPS (Zhang et al., 2018b) and FID (Heusel et al., 2017) respectively measure the pixel-wise reconstruction difference, perceptual difference, and distribution difference between the real and reconstructed images. The proposed masked noise encoder (Ours) consistently outperforms pSp, pSp† and the feature-style encoder. Note, pSp† is introduced by us, by training pSp with an additional discriminator and incorporating synthesized images for better initialization

Table 2 The effect of random noise masking on improving domain generalization via ISSA

Method	CS	ACDC	BDD	DarkZ
Baseline	**70.47**	41.48	45.66	15.25
ISSA w/o masking	69.68	44.63	46.45	17.36
ISSA w/- masking	69.48	**47.43**	**47.87**	**26.10**

Bold values indicate the best performance
We report the mean Intersection over Union (mIoU) of HRNet (Wang et al., 2021b) trained on Cityscapes at the resolution 256 × 512. BDD100K (BDD), ACDC, and Dark Zürich (DarkZ) represent different domain shifts from Cityscapes

encoder for data augmentation. Unfortunately, neither pSp† nor the Feature-Style encoder achieve satisfactory reconstruction quality. For instance, they both fail at capturing the red traffic light in Fig. 2. Using such images for data augmentation can confuse the semantic segmentation model, leading to performance degradation.

Ablation on the masking effect In Figs. 4 and 7, we visually observe that random masking offers a stronger perceivable style mixing effect compared to the model trained without masking. Next, we test the effect of masking on improving the domain generalization for the semantic segmentation task. In particular, we employ the encoder that is trained with and without masking to perform ISSA. In Table 2, while slightly degrading the source domain performance of the baseline model on Cityscapes, ISSA improves the domain generalization performance on BDD100K, ACDC and Dark Zürich. As ISSA with masked noise encoder is more effective at diversifying the training set and reducing the style-content correlation, it achieves more pronounced gains in Table 2, e.g., more than 10% improvement in mIoU from Cityscapes to Dark Zürich.

Table 3 Ablation on the mask patch size and masking ratio

Patch size	Ratio	MSE ↓	LPIPS ↓	FID ↓
2	25%	0.005	0.090	1.50
	50%	0.008	0.127	2.02
4	25%	**0.004**	**0.089**	**1.41**
	50%	0.009	0.129	2.01

Bold values indicate the best performance
The influence of patch size on the reconstruction is minor, while masking ratio is more important, i.e., higher masking ratio has negative impact

Table 4 Effect of noise map resolution on reconstruction quality

Noise scale	MSE ↓	LPIPS ↓	FID ↓
$4 \times 8 \sim 8 \times 16$	0.041	0.317	14.90
32×64	0.008	0.101	2.30

Experiments are done on Cityscapes, 128 × 256 resolution

Ablation on masking hyperparameters We conduct an ablation study on the mask patch size P and masking ratio ρ, shown in Table 3. We observe that the mask patch size is a relatively insensitive hyperparameter, while higher masking ratio results in noticeable degradation on the reconstruction quality. Empirically, the patch size $P = 4$ with a masking ratio $\rho = 25\%$ achieves the best reconstruction performance. Therefore, we use the encoder trained with this parameter combination for our data augmentation ISSA.

Ablation on the noise map resolution We investigate the effect of noise map size and experimentally observed that the reconstruction quality benefits the most from using the noise map at the intermediate feature space with one fourth of the input resolution. As shown in Table 4, using 32 × 64 noise, i.e., one fourth of the image resolution, achieves bet-

Fig. 8 Influence of the noise map resolution on style-mixing ability. Using higher resolution noise map, e.g., $H \times W$, leads to poor style-mixing ability. While too low resolution, e.g., $\frac{H}{16} \times \frac{W}{16}$, cannot reconstruct the scene faithfully

Table 5 Comparison of data augmentation for improving domain generalization, i.e., from Cityscapes (train) to ACDC (unseen)

	HRNet (Wang et al., 2021b)						SegFormer (Xie et al., 2021)					
Method	CS	Rain	Fog	Snow	Night	Avg	CS	Rain	Fog	Snow	Night	Avg
Baseline	70.47	44.15	58.68	44.20	18.90	41.48	67.90	50.22	60.52	48.86	28.56	47.04
ColorTransform	69.90	49.35	65.14	52.63	26.56	48.42	68.50	51.58	66.45	52.87	30.33	50.31
CutMix (Yun et al., 2019)	**72.68**	42.48	58.63	44.50	17.07	40.67	**69.23**	49.53	61.58	47.42	27.77	46.57
Hendrycks-Weather	69.25	**50.78**	60.82	38.34	22.82	43.19	67.41	54.02	64.74	49.57	28.50	49.21
Hendrycks-Digital	69.13	50.13	65.71	49.22	24.81	47.47	67.57	55.53	66.46	49.92	30.33	50.56
FDA (Yang & Soatto, 2020)	70.43	49.68	65.19	50.65	26.41	47.98	67.92	51.28	67.03	51.30	28.28	49.47
StyleMix (Hong et al., 2021)	57.40	40.59	49.11	39.14	19.34	37.04	65.30	53.54	63.86	49.98	28.93	49.08
ISSA (Ours)	70.30	50.62	**66.09**	**53.30**	**30.18**	**50.05**	67.52	**55.91**	**67.46**	**53.19**	**33.23**	**52.45**
Oracle	70.29	65.67	75.22	72.34	50.39	65.90	68.24	63.67	74.10	67.97	48.79	63.56

Bold values indicate the best performance

The mean Intersection over Union (mIoU) is reported on Cityscapes (CS), four individual scenarios of ACDC (Rain, Fog, Snow and Night) and the whole ACDC (Avg.). ColorTransform consists of various color transformations such as altering the contrast, brightness, saturation; luma flip and hue rotation. Hendrycks-Weather (Hendrycks & Dietterich, 2019) simulates weather conditions in a synthetic manner for data augmentation, and Hendrycks-Digital is composed of contrast, elastics transformation, pixelation and JPEG corruption. Oracle indicates the supervised training on both Cityscapes and ACDC, serving as an upper bound on ACDC for the other methods. Note, it is not supposed to be an upper bound on Cityscapes. Underline denotes worse results than the baseline on ACDC. ISSA performs the best and consistently improves the mIoU in all four scenarios of ACDC using both HRNet and SegFormer

Table 6 Comparison of data augmentation for improving domain generalization, i.e., from Cityscapes (train) to ACDC, BDD100K and Dark Zürich (unseen)

	HRNet(Wang et al., 2021b)				SegFormer (Xie et al., 2021)			
Method	CS	ACDC	BDD100K	Dark Zürich	CS	ACDC	BDD100K	Dark Zürich
Baseline	70.47	41.48	45.66	15.50	67.90	47.04	49.35	24.20
ColorTransform	69.90	48.42	50.22	24.13	68.50	50.31	51.09	25.04
CutMix (Yun et al., 2019)	**72.68**	40.67	45.57	15.34	**69.23**	46.57	48.93	22.98
Hendrycks-Weather	69.25	43.19	44.53	18.71	67.41	49.21	49.84	23.44
Hendrycks-Digital	69.13	47.47	47.60	22.32	67.57	50.56	51.11	25.11
FDA (Yang & Soatto, 2020)	70.43	47.98	48.74	22.46	67.92	49.47	50.47	22.45
StyleMix (Hong et al., 2021)	57.40	37.04	39.30	15.85	65.30	49.08	50.49	23.50
ISSA (Ours)	70.30	**50.05**	**50.29**	**27.24**	67.52	**52.45**	**51.92**	**27.39**

Bold values indicate the best performance

ISSA consistently outperforms the other data augmentation techniques across different datasets and network architectures, which is consistent with the Table 5

ter reconstruction quality than using lower resolution noise maps. Higher resolution noise map, e.g., full image resolution, in contrast, can be too expressive and encode nearly all perceivable details. This results in worse style mixing capability, as shown in Fig. 8. Therefore, we employ the intermediate noise map at one fourth of the input resolution in all of our experiments.

4.3 ISSA for Domain Generalization

Comparison with data augmentation methods Table 5 reports the mIoU scores of Cityscapes to ACDC domain generalization using two semantic segmentation models, i.e., HRNet (Wang et al., 2021b) and SegFormer (Xie et al., 2021). Qualitative visualization is illustrated in Fig. 9. ISSA is compared with three representative data augmentations methods, i.e., CutMix (Yun et al., 2019), Hendrycks's weather and digital corruptions (Hendrycks & Dietterich, 2019), and StyleMix (Hong et al., 2021). Remarkably, our ISSA is the top performing method, consistently improving mIoU in both models and across all four different scenarios of ACDC, i.e., rain, fog, snow and night. Compared to HRNet, SegFormer is more robust against the considered domain shifts.

In contrast to the others, CutMix mixes up the content rather than the style. It improves the in-distribution performance on Cityscapes, but this gain does not extend to domain generalization. Hendrycks's weather corruptions can be seen as the synthetic version of Cityscapes under the rain, fog, and snow weather conditions. While already mimicking ACDC at training, it can still degrade ACDC-Snow by

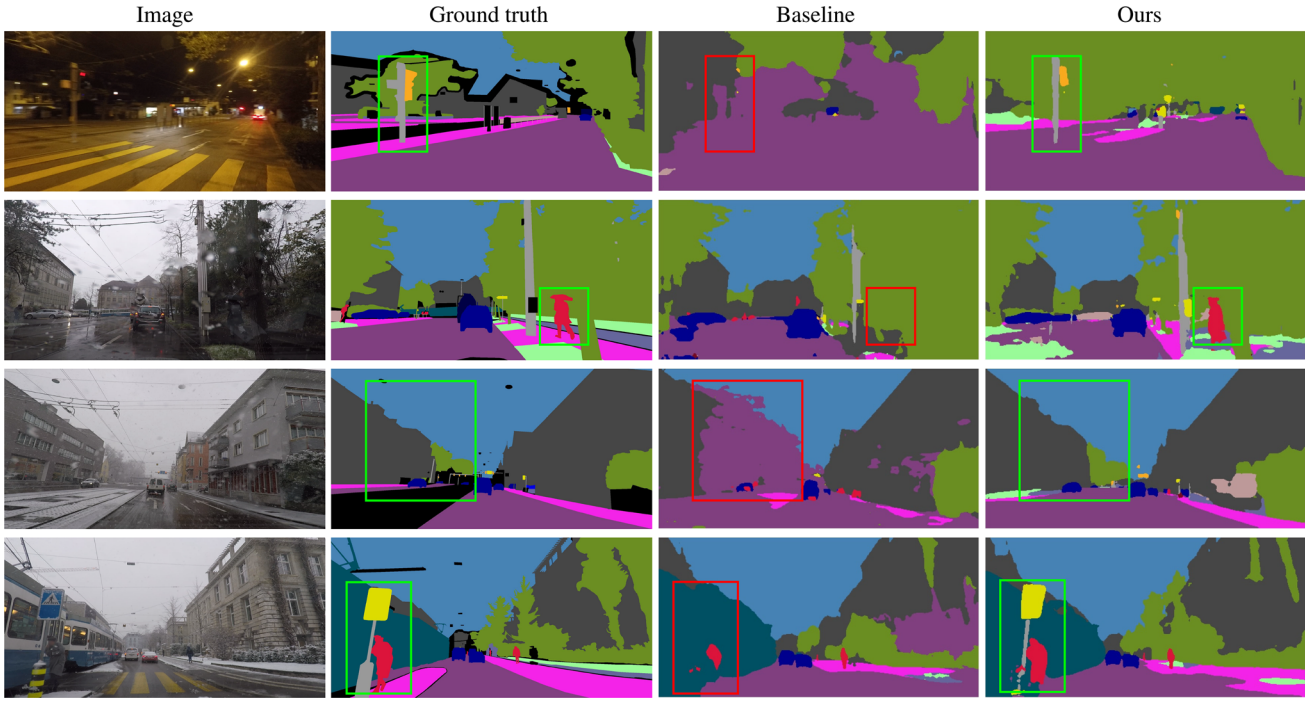

Fig. 9 Semantic segmentation results of Cityscapes to ACDC generalization using HRNet. The HRNet is trained on Cityscapes only. The segmenter trained with ISSA provides more reasonable prediction under adverse weather conditions

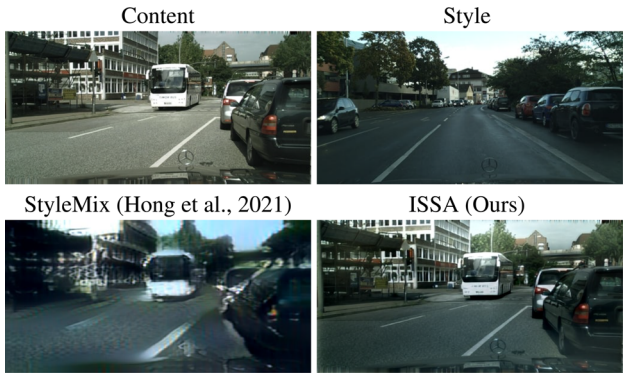

Fig. 10 Comparison of StyleMix (Hong et al., 2021) and ISSA. StyleMix has rather low fidelity, while ISSA can preserve more details

more than 5.8% in mIoU using HRNet. Among the four Hendrycks' corruption types (i.e., noise, blur, digital and weather), Hendrycks-Digital, consisting of contrast, elastics transformation, pixelation and JPEG, is the best-performing one, but still underperforms ISSA. StyleMix (Hong et al., 2021) also seeks to mix up styles. However, it does not work well for scene-centric datasets, such as Cityscapes. Its poor synthetic image quality (see Fig. 10) leads to the performance drop over the HRNet baseline in many cases, e.g., on Cityscapes to ACDC-Fog from 58.68 to 49.11% mIoU.

More evaluation on the generalization performance from Cityscapes to BDD100K and Dark Zürich is provided in Table 6, where the observation is consistent with Table 5 explained above. In addition to weather changes, we further compare different data augmentation methods under the more challenging day-to-night setting in Table 7. ISSA present consistent advantages over competing methods, which again justifies the effectiveness of ISSA on improving generalization performance.

Table 7 Comparison of data augmentation techniques for improving domain generalization using HRNet (Wang et al., 2021b), i.e., from BDD100K-Daytime to ACDC-Night and Dark Zürich

Method	BDD100K	ACDC-Night	DarkZürich
Baseline	52.97	23.52	23.63
CutMix	**54.03**	24.37	23.99
Weather	52.10	23.79	24.21
Digital	52.10	24.17	23.24
StyleMix	46.33	19.13	19.27
ISSA(Ours)	53.37	**25.93**	**26.55**

Bold values indicate the best performance
BDD100K-Daytime is a subset of BDD100K, which contains 2526 images in daytime under various weather conditions, but not in dawn/nighttime. Here, we evaluate the domain generalization with respect to day to night

Comparison with domain generalization techniques We further compare ISSA with two advanced feature space style mixing methods designed to improve domain generalization performance: MixStyle (Zhou et al., 2021) and DSU (Li et al.,

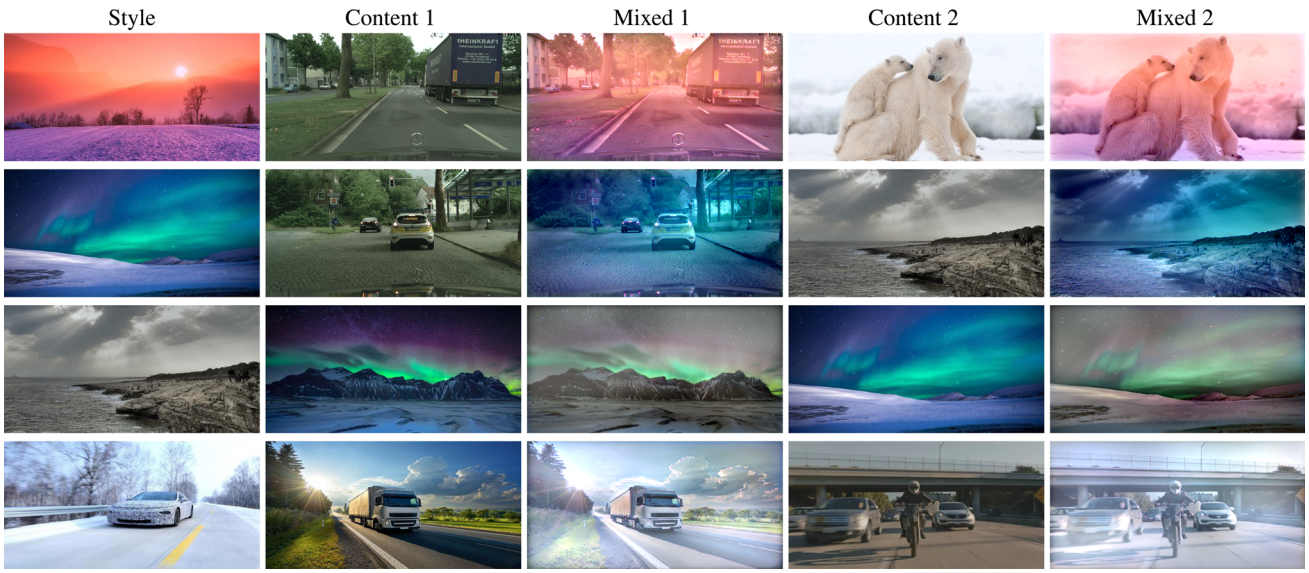

Fig. 11 Extra-source exemplar based style synthesis using web-crawled images, where the generator and encoder are only trained on Cityscapes. Except for the Content 1 image of the first 2 rows, all the others are web-crawled images

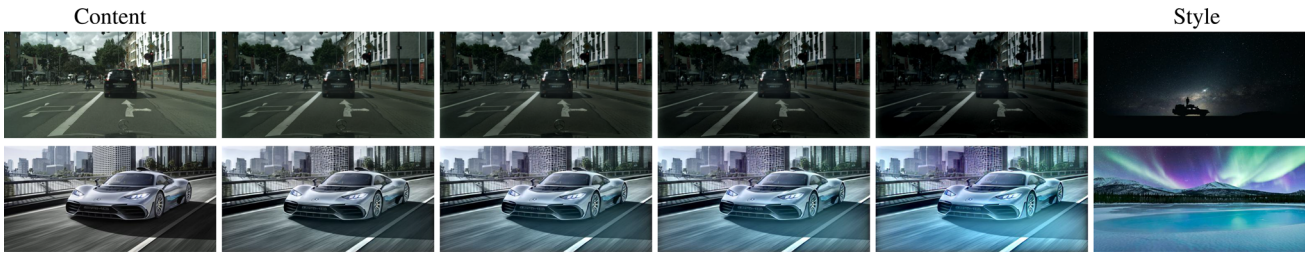

Fig. 12 Visualization of interpolation in the style latent space. As illustrated, we can control the style mixing strength and achieve a smooth transition on both trained Cityscapes and unseen web-crawled images

Table 8 Comparison with feature-level augmentation methods on domain generalization performance of Cityscapes as the source

Method	CS	ACDC	BDD	DarkZ
Baseline (Chen et al., 2017a)	61.73	30.86	34.30	11.62
MixStyle (Zhou et al., 2021)	59.01	36.97	36.27	9.38
DSU (Li et al., 2022)	59.59	38.31	35.53	12.29
ISSA (Ours)	**62.20**	**43.21**	**42.60**	**21.56**
MixStyle + ISSA	60.17	41.81	42.17	20.56
DSU + ISSA	60.20	43.31	42.24	24.63

Bold values indicate the best performance
Following DSU (Li et al., 2022), we conduct experiments using DeepLab v2 (Chen et al., 2017a) as the baseline for fair comparison

Table 9 Combination of ISSA and RobustNet (Choi et al., 2021)

Method	CS	ACDC	BDD	DarkZ
Baseline (Chen et al., 2018)	69.01	44.23	43.27	16.03
RobustNet (Choi et al., 2021)	**69.47**	47.25	46.94	20.11
+ ISSA	69.45	**47.55**	**48.44**	**23.09**
SHADE (Zhao et al., 2022)	**64.24**	47.30	46.44	25.37
+ ISSA	63.79	**47.64**	**47.76**	**25.58**

Bold values indicate the best performance
We adopt the experimental setting of RobustNet and use DeepLab v3+ (Chen et al., 2018) as the baseline. Our ISSA is complementary to RobustNet and further improves its generalization performance

2022). Both extract the style information at certain normalization layers of CNNs. MixStyle (Zhou et al., 2021) mixes up styles by linearly interpolating the feature statistics, i.e., mean and variance, of different images, while DSU (Li et al., 2022) models the feature statistics as a distribution and randomly draws samples from it.

We adopt the experimental setting of DSU with default hyperparameters, using DeepLab v2 (Chen et al., 2017a) segmentation network with ResNet101 backbone. Table 8 shows that ISSA outperforms both MixStyle and DSU by a large margin. We also observe that there is a slight performance drop on the source domain (i.e., CS) when applying DSU and MixStyle. As they operate at the feature-level, there is no guarantee that the semantic content stays unchanged after the

Table 10 Comparison with UDA methods on Cityscapes to ACDC generalization

Method	Network	Use target	mIoU
Baseline	DeepLabv2	–	30.9
BDL(Li et al., 2019b)		✓	32.7
CRST (Zou et al., 2019)		✓	32.8
AdaptSegNet(Tsai et al., 2018)		✓	33.4
SIM(Wang et al., 2020)		✓	34.6
MRNet(Zheng & Yang, 2021)		✓	36.1
ADVENT(Tsai et al., 2019)		✓	37.7
CLAN(Luo et al., 2019)		✓	39.0
FDA(Yang & Soatto, 2020)		✓	45.7
ISSA(Ours)		✗	43.2
DAFormer(Hoyer et al., 2022)	DAFormer	✓	55.4
ISSA(Ours)	SegFormer	✗	52.5

Remarkably, our domain generalization method (without access to the target domain, neither images nor labels), is on-par or better than unsupervised domain adaptation (UDA) methods, which requires knowledge of the target domain during training. Results of UDA methods are from(Sakaridis et al., 2021)

Table 11 Comparison on Cityscapes to ACDC generalization using ISSA with generator and encoder trained on Cityscapes (CS-G-E) and BDD100K (BDD-G-E), respectively

Method	CS	Rain	Fog	Snow	Night	Avg
Baseline	**70.5**	44.2	58.7	44.2	18.9	41.5
ISSA: CS-G-E	70.3	50.6	66.1	**53.3**	30.2	50.1
ISSA: BDD-G-E	70.3	**52.2**	**66.3**	52.2	**31.0**	**50.4**

Bold values indicate the best performance
Despite never seeing Cityscapes samples, ISSA with BDD-G-E is still highly effective

Table 12 Utilizing Landscape Pictures as extra-source exemplars for style augmentation, where the generator and encoder are only trained on Cityscapes (CS-G-E)

Method	CS	ACDC	BDD	DarkZ
Baseline	70.47	41.48	45.66	15.50
ISSA: CS-G-E	70.30	50.05	50.29	27.24
ESSA: CS-G-E	69.85	**50.87**	**51.42**	**29.06**

Bold values indicate the best performance
ESSA can further improve the generalization performance from Cityscapes to other unseen datasets

random perturbation of feature statistics. Thus, the changes in feature statistics might negatively affect the performance, as also indicated in (Li et al., 2022). Note that, in contrast, ISSA operates on the image space. Combining ISSA with MixStyle and DSU leads to a strong boost in performance of these methods.

Being model-agnostic, ISSA can be combined with other networks designed specifically for the domain generalization of semantic segmentation. To showcase its complementary nature, we add ISSA on top of two state-of-the-art domain generalization methods for semantic segmentation, RobustNet (Choi et al., 2021) and SHADE (Zhao et al., 2022). RobustNet proposed a novel instance whitening loss to selectively remove domain-specific style information. SHADE on the other hand aims to learn style-invariant representation and preserve knowledge from the pretrained backbone. Although color transformation has already been used for augmentation in both methods and SHADE additionally employs feature-level style augmentation, ISSA can introduce more natural style shifts, thus is able to bring further improvements. Table 9 verifies the effectiveness of ISSA, which brings extra gains for RobustNet and SHADE. For RobustNet, the performance of the challenging day to night scenario, i.e., Cityscapes to Dark Zürich is boosted from 20.11 to 23.09% in mIoU.

Comparison with unsupervised domain adaptation methods We compare our method with multiple unsupervised domain adaptation (UDA) techniques, which not only have access to the source domain, but also use extra unlabeled samples of the target domain. The quantitative comparison of Cityscapes to ACDC adaptation/generalization is shown in Table 10. Our method has presented competitive performance, even without using images from the target domain.

5 Plug-n-Play Ability of the Exemplar-Based Style Synthesis Pipeline

In Sect. 4.3, we have focused on ISSA for improved domain generalization. Next, we investigate the plug-n-play ability of our exemplar-based style pipeline, which enables ESSA. Specifically, the generator and masked noise encoder which are trained on one dataset can be directly used for mixing styles from other datasets, thus avoiding retraining or fine-tuning the models. This ability is valuable in two perspectives: (1) harnessing external data for improved domain generalization via ESSA; and (2) saving computationally complexity. Compared to other data augmentation techniques such as CutMix (Yun et al., 2019), Hendrycks corruption (Hendrycks & Dietterich, 2019), our style synthesis requires training GAN and an encoder, which could take considerable computational resources. Therefore, it is of practical interest if the trained models can be readily useable for novel domains.

ISSA using arbitrary encoders Favorably, thanks to the plug-n-play ability of the synthesis pipeline, we observe that ISSA can still be effective even when encoder and generator are

Fig. 13 Visual examples of stylized data by transferring style from one unannotated ACDC sample (target domain) to Cityscapes (source domain). Best view in color (Color figure online)

trained on a different dataset of a similar task, and re-training is not required. Note that here the source is with respect to the segmenter training for domain generalization, not the encoder training. As shown in Table 11, when training the segmenter on Cityscapes using ISSA, we can directly use generator and encoder trained on BDD100K without fine-tuning. Even though the models have not seen any samples of Cityscapes, they can still reconstruct and augment styles within Cityscapes, and the effectiveness of ISSA is not compromised. This implies that, once the generator and encoder are trained on one dataset, they are also straightforwardly applicable for augmenting novel datasets.

Extra-source exemplar based style synthesis Furthermore, we exploit the usage of extra-source data as the style exemplar. Visual examples in Fig. 11 showcase the plug-n-play style-mixing ability of our encoder on web-crawled images, where the model is only trained on Cityscapes. It can be observed that the style of unseen images can still be successfully transferred to the content images, which grants us the opportunity to further utilize images on the web to enhance the effectiveness of style augmentation beyond intra-source styles. Also, we illustrate the interpolation capability in the style latent space on both trained Cityscapes and unseen web-crawled image. This property enables more control on the style mixing strength.

To further explore the usage of images on the web, we take Landscape Pictures[1] dataset as the extra-source exemplars for style augmentation. Table 12 justifies that by exploiting additional image styles, ESSA can further improve the generalization performance of ISSA on unseen target domains.

[1] https://www.kaggle.com/datasets/arnaud58/landscape-pictures?resource=download

6 Stylized Proxy Validation Set Synthesis

Beyond the usage of data augmentation for network training, we further explore if our exemplar-based style synthesis pipeline can be used to assess the generalization capability of semantic segmentation models for both source and target domain without extra data annotation effort. Prior work (Zhang et al., 2021b) has used conditional GAN synthesized samples to predict generalization performance of image classifiers in the source domain. However, it remains unclear how to evaluate the generalization performance on unseen domains, and apply it on dense prediction tasks. Given the fact that our masked noise encoder can transfer styles even from novel domains, we utilize this attractive property to generate a stylized proxy validation set, i.e., combining styles from the target domain with the contents from the source domain training samples. For getting their styles, exemplars from the target domain do not need to be labelled. The existing ground-truth label maps of the training samples in the source domain are reused as the ground-truth annotations of the stylized proxy validation set. Visual examples of transferring ACDC style using one sample from each weather condition are provided in Fig. 13.

Experimental Setup We investigate the generalization performance of 95 semantic segmentation models trained on Cityscapes, where 54 models are obtained from MMSeg-

mentation (Contributors, 2020) model zoo and the others are trained by ourselves. The models cover both CNN-based architectures, e.g., HRNet (Wang et al., 2021b), DeepLab (Chen et al., 2017b), DANet (Fu et al., 2019), and transformer-based model, e.g., SegFormer (Xie et al., 2021), SETR (Zheng et al., 2021). Besides, the models are trained using different strategies, e.g., various learning rate schedule, cropping size and data augmentation. We consider generalization performance on both source and target domain for the correlation study. Specifically, we use the Cityscapes validation set as the source test set, ACDC and BDD100K validation sets as the target test data. To verify the generalization performance on the source domain, we apply intra-source style augmentation on the Cityscapes training set and use it as the proxy validation set. For the verification of target domain generalization performance, we build a proxy set by transferring styles from the corresponding target test dataset. Further, we study the correlation between the real test performance and performance on the proxy data.

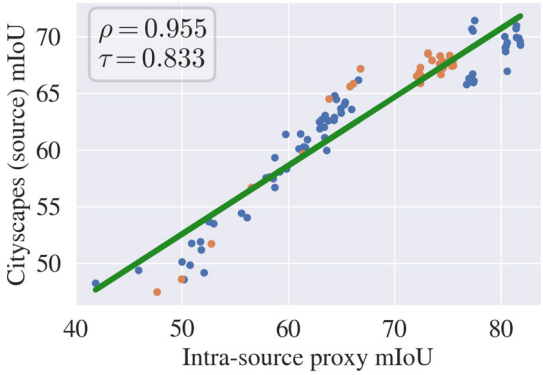

Fig. 14 Correlation between real Cityscapes test performance and intra-source style augmented proxy performance for 95 models. Spearman's Rank Correlation coefficient (ρ) and Kendall Rank Correlation Coefficient (τ) are computed to quantitatively measure correlation strength. Blue and orange dots represent CNN- and transformer-based backbones, respectively. We observe that there is a strong correlation between the real test mIoU and proxy mIoU (Color figure online)

Correlation Metrics We compute Spearman's Rank Correlation coefficient (ρ) and Kendall Rank Correlation Coefficient (τ) to quantitatively measure the correlation strength. The value of the correlation coefficient varies from $[-1, 1]$. A value closer to ± 1 indicates strong positive/negative association between the two variables. As the coefficient goes towards 0, the association becomes looser. Both correlation coefficients are non-parametric, i.e., no strict assumptions on the data distribution, and the assessment is based on the ranking of the data.

Observations In Fig. 14, we show the correlation of performance on the intra-source style augmented proxy set and real Cityscapes test set across different network architectures. We clearly observe a strong correlation ($\rho > 0.95$), indicating that ISSA proxy set can serve as a good indicator for generalization in the source domain.

Furthermore, we report the correlation results of target domain generalization on two datasets, i.e., ACDC and BDD100K in each row of Fig. 15. We compare three different choices of the proxy set in each column, namely the original Cityscapes validation set, intra-source style augmented Cityscapes validation set and target data style augmented validation set. Blue and orange dots represent CNN- and transformer-based backbones, respectively. Quantitatively, the correlation coefficients of Fig. 15a, d are rather low. Also from Fig. 15a, some blue points in the upper right corner has stronger performance on Cityscapes validation set compared to the orange points, but worse on ACDC test data. This suggests that evaluation of the original Cityscapes (source) validation set cannot properly reflect the generalization performance on the target domain. Therefore, this raises the concern that by following the traditional way, selecting the best model based on the source validation performance could be problematic when the deploying environment involves data of unknown target domains. By applying intra-source style augmentation on the Cityscapes validation set, the correlation coefficient has been improved (see Fig. 15b, e). We hypothesize that style mixing results in better data coverage and thus can better represent model's generalization ability under style shifts. Furthermore, whenever it is possible to have access to images of the target domain, even though without annotation, we can utilize styles of the unlabeled target data and achieve the strongest correlation in Fig. 15c, f. In addition to the correlation metric, in general models have higher mIoU on the Cityscapes validation set, compared with the intra-source style and target domain style augmented proxy set. And the mIoU range on the intra-source proxy set is closer to the one of using target styles, which also justifies our hypothesis above.

Besides, we also observe an interesting phenomenon from Fig. 15: all transformer-based models (orange dots) are above the linear fit. This suggests that transformer-based models present better generalization ability under natural shifts compared with CNN-based models (blue dots). This is consistent with the acknowledgement on transformers from prior works (Naseer et al., 2021; Bai et al., 2021; Zhang et al., 2022).

To sum up, we present a new use case of proposed exemplar-based style synthesis pipeline, and demonstrate that stylized samples can be used as a proxy validation set and a strong indicator for model's generalization capability without introducing additional annotation efforts. Based on this observation, we can better utilize existing annotated data together with our exemplar-based style synthesis pipeline, to select models in practice especially when deployment in an

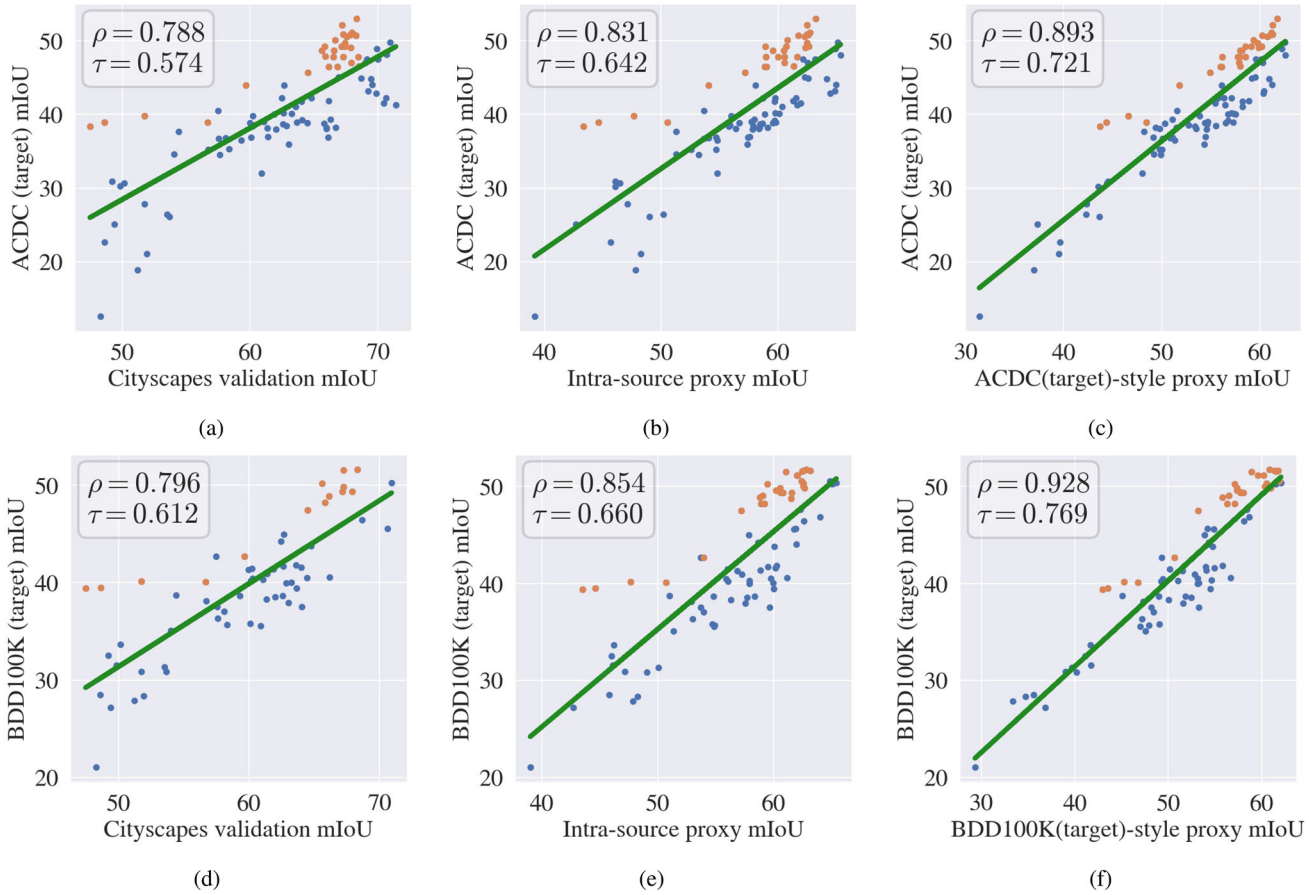

Fig. 15 Correlation between test performance and proxy performance for 95 models. We compute Spearman's Rank Correlation coefficient (ρ) and Kendall Rank Correlation Coefficient (τ) to quantitatively measure correlation strength. Blue and orange dots represent CNN- and transformer-based backbones, respectively. In each row, we investigate the correlation between the real test performance, i.e., mIoU of ACDC and BDD100K, and mIoU of different proxy sets. We observe that Fig. 15c, f achieve the strongest correlation for each scenario, which indicates that it is beneficial to build a proper proxy set using styles of the corresponding test dataset (Color figure online)

open-world environment, where unknown target data commonly exists.

7 Conclusion and Discussions

In this paper, we propose a GAN inversion based style synthesis pipeline for domain generalization in semantic segmentation. The key enabler for our pipeline is the masked noise encoder, which is capable of preserving fine-grained content details and allows style mixing between images without affecting the semantic content. In particular, we employ intra-source style augmentation (ISSA) for learning domain generalized semantic segmentation using restricted training data from a single source domain. Extensive experimental results verify the effectiveness of ISSA on domain generalization across different datasets and network architectures. We further demonstrate the plug-n-play ability of the proposed pipeline. Without requiring retraining the encoder and generator, our model can be used directly on extra-source exemplars such as web-crawled images, enabling extra-source style augmentation (ESSA). It also opens up applications beyond data augmentation for improved domain generalization. Specifically, we show that the intra- & extra-source exemplar-based style synthesis pipeline can be used for creating proxy validation sets to compare the generalization capability of diverse models on both the source and target domain without extra data annotation effort.

Limitation and future work One limitation of ISSA is that our style mixing is a global transformation, which cannot specifically alter the style of local objects, e.g., adjusting vehicle color from red to black, though when changing the image globally, local areas are inevitably modified. Also compared to simple data augmentation such as color transformation, our pipeline requires higher computational complexity for training. It takes around 7 days to train the masked noise encoder on 256×512 resolution using 2 GPUs.

A similar amount of time is required for the StyleGAN2 training. Nonetheless, for data augmentation, it only concerns the inference time of our encoder, which is much faster, i.e., 0.1 s, compared to optimization based methods such as PTI (Roich et al., 2022) that takes 55.7 s per image.

In the future, it is challenging yet interesting to extend our work with more flexible local editing. Our proposed intra- & extra-source exemplar-based style synthesis is a global transformation, which cannot specifically alter the style of local objects, e.g., adjusting vehicle color from red to black, though when changing the image globally, local areas are inevitably modified. One potential direction is by exploiting the pre-trained language-vision model, such as CLIP (Radford et al., 2021). We can synthesize styles conditioned on text rather than an image. For instance, by providing a text condition "snowy road", ideally we would want to obtain an image where there is snow on the road and other semantic classes remain unchanged. Recent works (Bar-Tal et al., 2022; Hertz et al., 2022; Kawar et al., 2023) studied local editing conditioned on text. However, CLIP exhibits a strong bias (Bar-Tal et al., 2022) and may generate undesirable results, and the edited image may suffer from insufficient alignment with the other parts of the image. Overall, there is still large room for improvement on synthesizing images with more controls on both style and content.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability The datasets analysed during the current study are available at Cityscapes, ACDC, BDD100K, Dark Zürich, Landscape Pictures repository, respectively.

References

Abdal R, Qin Y, Wonka P (2019) Image2stylegan: How to embed images into the stylegan latent space? In: ICCV, pp 4432–4441

Abdal R, Qin Y, Wonka P (2020) Image2stylegan++: How to edit the embedded images? In: CVPR, pp 8296–8305

Alaluf Y, Tov O, Mokady R, Gal R, Bermano A (2022) Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In: CVPR, pp 18511–18521

Bai, Y., Mei, J., Yuille, A. L., & Xie, C. (2021). Are transformers more robust than cnns? *NeurIPS, 34*, 26831–26843.

Balaji Y, Sankaranarayanan S, Chellappa R (2018) Metareg: Towards domain generalization using meta-regularization. In: NeurIPS, vol 31

Bar-Tal O, Ofri-Amar D, Fridman R, Kasten Y, Dekel T (2022) Text2live: Text-driven layered image and video editing. In: ECCV, pp 707–723

Bartz C, Bethge J, Yang H, Meinel C (2021) One model to reconstruct them all: A novel way to use the stochastic noise in StyleGAN. In: BMVC, p 89

Bin Y, Cao X, Chen X, Ge Y, Tai Y, Wang C, Li J, Huang F, Gao C, Sang N (2020) Adversarial semantic data augmentation for human pose estimation. In: ECCV, pp 606–622

Burton S, Gauerhof L, Heinzemann C (2017) Making the case for safety of machine learning in highly automated driving. In: SAFECOMP, pp 5–16

Caesar H, Uijlings J, Ferrari V (2018) Coco-stuff: Thing and stuff classes in context. In: CVPR, pp 1209–1218

Chai L, Zhu JY, Shechtman E, Isola P, Zhang R (2021) Ensembling with deep generative views. In: CVPR, pp 14997–15007

Chen C, Li J, Han X, Liu X, Yu Y (2022) Compound domain generalization via meta-knowledge encoding. In: CVPR, pp 7119–7129

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI, 40*(4), 834–848. https://doi.org/10.1109/TPAMI.2017.2699184

Chen LC, Papandreou G, Schroff F, Adam H (2017b) Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587

Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV, pp 801–818

Choi S, Jung S, Yun H, Kim JT, Kim S, Choo J (2021) RobustNet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In: CVPR, pp 11580–11590

Collins E, Bala R, Price B, Susstrunk S (2020) Editing in style: Uncovering the local semantics of gans. In: CVPR, pp 5771–5780

Contributors M (2020) MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation

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, pp 3213–3223

Creswell, A., & Bharath, A. A. (2019). Inverting the generator of a generative adversarial network. *TNNLS, 30*(7), 1967–1974. https://doi.org/10.1109/TNNLS.2018.2875194

Dabouei A, Soleymani S, Taherkhani F, Nasrabadi NM (2021) Supermix: Supervising the mixing data augmentation. In: CVPR, pp 13794–13803

Deng Z, Ding F, Dwork C, Hong R, Parmigiani G, Patil P, Sur P (2020) Representation via representations: Domain generalization via adversarially learned invariant representations. arXiv preprint arXiv:2006.11478

DeVries T, Taylor GW (2017) Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552

Dinh TM, Tran AT, Nguyen R, Hua BS (2022) Hyperinverter: Improving stylegan inversion via hypernetwork. In: CVPR, pp 11389–11398

D'Innocente A, Caputo B (2018) Domain generalization with domain-specific aggregation modules. In: GCPR, pp 187–198

Fu J, Liu J, Tian H, Li Y, Bao Y, Fang Z, Lu H (2019) Dual attention network for scene segmentation. In: CVPR, pp 3146–3154

Gatys LA, Ecker AS, Bethge M (2016) Image style transfer using convolutional neural networks. In: CVPR, pp 2414–2423

Golhar M, Bobrow TL, Ngamruengphong S, Durr NJ (2022) GAN Inversion for Data Augmentation to Improve Colonoscopy Lesion Classification. arXiv preprint arXiv:2205.02840

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: NeurIPS, pp 3481–3490

Gu J, Shen Y, Zhou B (2020) Image processing using multi-code gan prior. In: CVPR, pp 3012–3021

He K, Chen X, Xie S, Li Y, Dollár P, Girshick R (2022) Masked autoencoders are scalable vision learners. In: CVPR, pp 16000–16009

Hendrycks D, Dietterich T (2019) Benchmarking neural network robustness to common corruptions and perturbations. In: ICLR https://openreview.net/forum?id=HJz6tiCqYm

Hendrycks D, Mu N, Cubuk ED, Zoph B, Gilmer J, Lakshminarayanan B (2020) AugMix: A simple method to improve robustness and uncertainty under data shift. In: ICLR https://openreview.net/forum?id=S1gmrxHFvB

Hertz A, Mokady R, Tenenbaum J, Aberman K, Pritch Y, Cohen-Or D (2022) Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626

Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S (2017) GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS, vol 30

Hoffman J, Tzeng E, Park T, Zhu JY, Isola P, Saenko K, Efros A, Darrell T (2018) Cycada: Cycle-consistent adversarial domain adaptation. In: ICML, pp 1989–1998

Hong M, Choi J, Kim G (2021) StyleMix: Separating content and style for enhanced data augmentation. In: CVPR, pp 6438–6447

Hoyer L, Dai D, Van Gool L (2022) Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In: CVPR, pp 9924–9935

Hu S, Zhang K, Chen Z, Chan L (2020) Domain generalization via multidomain discriminant analysis. In: UAI, pp 292–302

Hu X (2022) Invgan: Invertible gans. In: GCPR, pp 3–19

Huang J, Guan D, Xiao A, Lu S (2021) Fsdr: Frequency space domain randomization for domain generalization. In: CVPR, pp 6891–6902

Huang X, Belongie S (2017) Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. In: ICCV, pp 1501–1510

Jia Y, Zhang J, Shan S, Chen X (2020) Single-side domain generalization for face anti-spoofing. In: CVPR, pp 8484–8493, 10.1109/CVPR42600.2020.00851

Jiang, L., Dai, B., Wu, W., & Loy, C. C. (2021). Deceive d: Adaptive pseudo augmentation for gan training with limited data. *NeurIPS, 34*, 21655–21667.

Jin X, Lan C, Zeng W, Chen Z (2020) Feature alignment and restoration for domain generalization and adaptation. arXiv preprint arXiv:2006.12009

Kang K, Kim S, Cho S (2021) Gan inversion for out-of-range images with geometric transformations. In: ICCV, pp 13941–13949

Karras T, Aila T, Laine S, Lehtinen J (2018) Progressive Growing of GANs for Improved Quality, Stability, and Variation. In: ICLR https://openreview.net/forum?id=Hk99zCeAb

Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. In: CVPR, pp 4401–4410

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., & Aila, T. (2020). Training generative adversarial networks with limited data. *NeurIPS, 33*, 12104–12114.

Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T (2020b) Analyzing and improving the image quality of stylegan. In: CVPR, pp 8110–8119

Kawar B, Zada S, Lang O, Tov O, Chang H, Dekel T, Mosseri I, Irani M (2023) Imagic: Text-based real image editing with diffusion models. In: CVPR, pp 6007–6017

Khirodkar R, Yoo D, Kitani K (2019) Domain randomization for scene-specific car detection and pose estimation. In: WACV, pp 1932–1940

Kim J, Lee J, Park J, Min D, Sohn K (2022) Pin the memory: Learning to generalize semantic segmentation. In: CVPR, pp 4350–4360

Kim N, Son T, Pahk J, Lan C, Zeng W, Kwak S (2023) WEDGE: web-image assisted domain generalization for semantic segmentation. In: ICRA, pp 9281–9288

Lee K, Kim S, Kwak S (2022a) Cross-domain ensemble distillation for domain generalization. In: ECCV, pp 1–20

Lee S, Seong H, Lee S, Kim E (2022b) WildNet: Learning domain generalized semantic segmentation from the wild. In: CVPR, pp 9936–9946

Li D, Yang Y, Song YZ, Hospedales T (2018a) Learning to generalize: Meta-learning for domain generalization. In: AAAI, vol 32

Li D, Zhang J, Yang Y, Liu C, Song YZ, Hospedales T (2019a) Episodic training for domain generalization. In: ICCV, pp 1446–1455

Li H, Pan SJ, Wang S, Kot AC (2018b) Domain generalization with adversarial feature learning. In: CVPR, pp 5400–5409

Li, H., Wang, Y., Wan, R., Wang, S., Li, T. Q., & Kot, A. (2020). Domain generalization for medical imaging classification with linear-dependency regularization. *NeurIPS, 33*, 3118–3129.

Li X, Dai Y, Ge Y, Liu J, Shan Y, DUAN L (2022) Uncertainty Modeling for Out-of-Distribution Generalization. In: ICLR https://openreview.net/forum?id=6HN7LHyzGgC

Li Y, Yuan L, Vasconcelos N (2019b) Bidirectional learning for domain adaptation of semantic segmentation. In: CVPR, pp 6936–6945

Li Y, Zhang D, Keuper M, Khoreva A (2023) Intra-source style augmentation for improved domain generalization. In: WACV, pp 509–519

Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: CVPR, pp 2117–2125

Luo Y, Zheng L, Guan T, Yu J, Yang Y (2019) Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In: CVPR, pp 2507–2516

Mancini M, Bulo SR, Caputo B, Ricci E (2018) Best sources forward: Domain generalization through source-specific nets. In: ICIP, pp 1353–1357

Moon SJ, Park GM (2022) Interestyle: Encoding an interest region for robust stylegan inversion. In: ECCV, pp 460–476

Naseer, M. M., Ranasinghe, K., Khan, S. H., Hayat, M., Shahbaz Khan, F., & Yang, M. H. (2021). Intriguing properties of vision transformers. *NeurIPS, 34*, 23296–23308.

Nguyen DT, Tran CT, Nguyen TT, Hoang CB, Luu VP, Nguyen BN, Cheong PI (2021) Data augmentation for small face datasets and face verification by generative adversarial networks inversion. In: KSE, pp 1–6

Ouyang, C., Chen, C., Li, S., Li, Z., Qin, C., Bai, W., & Rueckert, D. (2022). Causality-inspired single-source domain generalization for medical image segmentation. *IEEE Transactions on Medical Imaging, 42*(4), 1095–1106. https://doi.org/10.1109/TMI.2022.3224067

Pan, X., Zhan, X., Dai, B., Lin, D., Loy, C. C., & Luo, P. (2022). Exploiting deep generative prior for versatile image restoration and manipulation. *TPAMI, 44*(11), 7474–7489. https://doi.org/10.1109/TPAMI.2021.3115428

Peng X, Tang Z, Yang F, Feris RS, Metaxas D (2018) Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In: CVPR, pp 2226–2234

Qiao F, Zhao L, Peng X (2020) Learning to learn single domain generalization. In: CVPR, pp 12556–12565, 10.1109/CVPR42600.2020.01257

Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, et al. (2021) Learning transferable visual models from natural language supervision. In: ICML, pp 8748–8763

Rahman MM, Fookes C, Baktashmotlagh M, Sridharan S (2020) Correlation-aware adversarial domain adaptation and generalization. PR 100:107124, 10.1016/j.patcog.2019.107124

Richardson E, Alaluf Y, Patashnik O, Nitzan Y, Azar Y, Shapiro S, Cohen-Or D (2021) Encoding in style: a stylegan encoder for image-to-image translation. In: CVPR, pp 2287–2296

Roich, D., Mokady, R., Bermano, A. H., & Cohen-Or, D. (2022). Pivotal tuning for latent-based editing of real images. *TOG, 42*(1), 1–13.

Sakaridis C, Dai D, Gool LV (2019) Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In: ICCV, pp 7374–7383

Sakaridis C, Dai D, Van Gool L (2021) Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In: ICCV, pp 10765–10775

Shafaei S, Kugele S, Osman MH, Knoll A (2018) Uncertainty in machine learning: A safety perspective on autonomous driving. In: SAFECOMP, pp 458–464

Shao R, Lan X, Li J, Yuen PC (2019) Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In: CVPR, pp 10023–10031, 10.1109/CVPR.2019.01026

Somavarapu N, Ma CY, Kira Z (2020) Frustratingly simple domain generalization via image stylization. arXiv preprint arXiv:2006.11207

Song H, Du Y, Xiang T, Dong J, Qin J, He S (2022) Editing out-of-domain gan inversion via differential activations. In: ECCV, pp 1–17

Šubrtová A, Futschik D, Čech J, Lukáč M, Shechtman E, Sỳkora D (2022) Chunkygan: Real image inversion via segments. In: ECCV, pp 189–204

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., & Schmidt, L. (2020). Measuring robustness to natural distribution shifts in image classification. *NeurIPS, 33*, 18583–18599.

Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., & Cohen-Or, D. (2021). Designing an encoder for stylegan image manipulation. *TOG, 40*(4), 1–14. https://doi.org/10.1145/3450626.3459838

Tsai YH, Hung WC, Schulter S, Sohn K, Yang MH, Chandraker M (2018) Learning to adapt structured output space for semantic segmentation. In: CVPR, pp 7472–7481

Tsai YH, Sohn K, Schulter S, Chandraker M (2019) Domain adaptation for structured output via discriminative patch representations. In: ICCV, pp 1456–1465 https://doi.org/10.1109/ICCV.2019.00154

Verma V, Lamb A, Beckham C, Najafi A, Mitliagkas I, Lopez-Paz D, Bengio Y (2019) Manifold mixup: Better representations by interpolating hidden states. In: ICML, pp 6438–6447

Voreiter C, Burnel JC, Lassalle P, Spigai M, Hugues R, Courty N (2020) A cycle gan approach for heterogeneous domain adaptation in land use classification. In: IGARSS, pp 1961–1964

Wan C, Shen X, Zhang Y, Yin Z, Tian X, Gao F, Huang J, Hua XS (2022) Meta convolutional neural networks for single domain generalization. In: CVPR, pp 4682–4691

Wang J, Jin S, Liu W, Liu W, Qian C, Luo P (2021a) When human pose estimation meets robustness: Adversarial algorithms and benchmarks. In: CVPR, pp 11855–11864

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al. (2021). Deep high-resolution representation learning for visual recognition. *TPAMI*. https://doi.org/10.1109/TPAMI.2020.2983686

Wang Z, Yu M, Wei Y, Feris R, Xiong J, Hwu Wm, Huang TS, Shi H (2020) Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In: CVPR, pp 12635–12644

Wang Z, Luo Y, Qiu R, Huang Z, Baktashmotlagh M (2021c) Learning to diversify for single domain generalization. In: ICCV, pp 834–843, 10.1109/ICCV48922.2021.00087

Wei, T., Chen, D., Zhou, W., Liao, J., Zhang, W., Yuan, L., Hua, G., & Yu, N. (2022). E2Style: Improve the efficiency and effectiveness of stylegan inversion. *TIP, 31*, 3267–3280. https://doi.org/10.1109/TIP.2022.3167305

Wu G, Gong S (2021) Collaborative optimization and aggregation for decentralized domain generalization and adaptation. In: ICCV, pp 6484–6493

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, 43*, 3349–3364.

Xie Z, Zhang Z, Cao Y, Lin Y, Bao J, Yao Z, Dai Q, Hu H (2022) SimMIM: A simple framework for masked image modeling. In: CVPR, pp 9653–9663

Yang Y, Soatto S (2020) FDA: Fourier domain adaptation for semantic segmentation. In: CVPR, pp 4085–4095

Yao X, Newson A, Gousseau Y, Hellier P (2022) Feature-Style Encoder for Style-Based GAN Inversion. arXiv preprint arXiv:2202.02183

Yu F, Seff A, Zhang Y, Song S, Funkhouser T, Xiao J (2015) Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365

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, pp 2636–2645

Yun S, Han D, Oh SJ, Chun S, Choe J, Yoo Y (2019) Cutmix: Regularization strategy to train strong classifiers with localizable features. In: ICCV, pp 6023–6032

Zhang C, Zhang M, Zhang S, Jin D, Zhou Q, Cai Z, Zhao H, Liu X, Liu Z (2022) Delving deep into the generalization of vision transformers under distribution shifts. In: CVPR, pp 7277–7286

Zhang H, Cisse M, Dauphin YN, Lopez-Paz D (2018a) mixup: Beyond Empirical Risk Minimization. In: ICLR https://openreview.net/forum?id=r1Ddp1-Rb

Zhang R, Isola P, Efros AA, Shechtman E, Wang O (2018b) The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR, pp 586–595

Zhang Y, Carballo A, Yang H, Takeda K (2021a) Autonomous Driving in Adverse Weather Conditions: A Survey. arXiv preprint arXiv:2112.08936

Zhang Y, Gupta A, Saunshi N, Arora S (2021b) On predicting generalization using gans. In: ICLR

Zhao Y, Zhong Z, Yang F, Luo Z, Lin Y, Li S, Sebe N (2021) Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification. In: CVPR, pp 6277–6286

Zhao Y, Zhong Z, Zhao N, Sebe N, Lee GH (2022) Style-hallucinated dual consistency learning for domain generalized semantic segmentation. In: ECCV, pp 535–552

Zheng S, Lu J, Zhao H, Zhu X, Luo Z, Wang Y, Fu Y, Feng J, Xiang T, Torr PH, et al. (2021) Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: CVPR, pp 6881–6890

Zheng, Z., & Yang, Y. (2021). Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *IJCV, 129*(4), 1106–1120. https://doi.org/10.1007/s11263-020-01395-y

Zhou F, Jiang Z, Shui C, Wang B, Chaib-draa B (2020) Domain generalization with optimal transport and metric learning. arXiv preprint arXiv:2007.10573

Zhou K, Yang Y, Qiao Y, Xiang T (2021) Domain generalization with mixstyle. In: ICLR https://openreview.net/forum?id=6xHJ37MVxxp

Zhu J, Shen Y, Zhao D, Zhou B (2020) In-domain gan inversion for real image editing. In: ECCV, pp 592–608

Zou Y, Yu Z, Liu X, Kumar B, Wang J (2019) Confidence regularized self-training. In: ICCV, pp 5982–5991