Intra-Source Style Augmentation for Improved Domain Generalization

Yumeng Li[1,2] Dan Zhang[1,4] Margret Keuper[2,3] Anna Khoreva[1,4]
[1]Bosch Center for AI [2] University of Siegen [3]MPI for Informatics [4]University of Tübingen
{yumeng.li, dan.zhang2, anna.khoreva}@de.bosch.com margret.keuper@uni-siegen.de

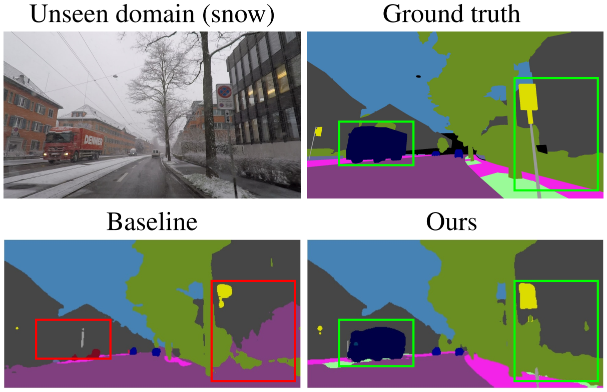
Abstract

*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 intra-source style augmentation (*ISSA*) method 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,* ISSA *effectively increases the diversity of training data and reduces spurious correlation. As a result, we achieve up to* 11.3% ***mIoU*** *improvements on driving-scene semantic segmentation under domain shifts, e.g., adverse weather conditions.*

1. Introduction

The varying environment with potentially diverse illumination and adverse weather conditions makes the deployment of deep learning models challenging in an open-world [20, 31]. 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 [3, 21].

Diversity of the training data is considered to play an important role for domain generalization, including natural distribution shifts [22]. However, for pixel-level prediction tasks such as semantic segmentation, collecting diverse training data involves a tedious and costly annotation process [4]. Therefore, improving generalization from a *single source domain* is exceptionally compelling, particularly for semantic segmentation.



Unseen domain (snow)

Ground truth

Baseline

Ours

Figure 1. Semantic segmentation results of HRNet [23] on unseen domain (snow), trained on Cityscapes [5] and tested on ACDC [20]. The model trained with our ISSA can successfully segment the truck, while the baseline model fails completely.

One pragmatic way to improve data diversity is by applying data augmentation. 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 [6], and CutMix [29]. However, they are ineffective on natural domain shifts as reported in [22]. Style augmentation, on the other hand, only modifies the style - the non-semantic appearance such as texture and color of the image [8] - while preserving the semantic content. Hendrycks corruptions [10] 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 intra-source style augmentation (ISSA) strategy for semantic segmentation, aiming to improve the style diversity in the training set without extra labeling effort or using extra data sources.

Our augmentation technique is based on the inversion of StyleGAN2 [16], 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 to encode a given image to latent variables, and thus facilitates faithful reconstruction with style mixing capability. To realize ISSA, we learn to separate semantic

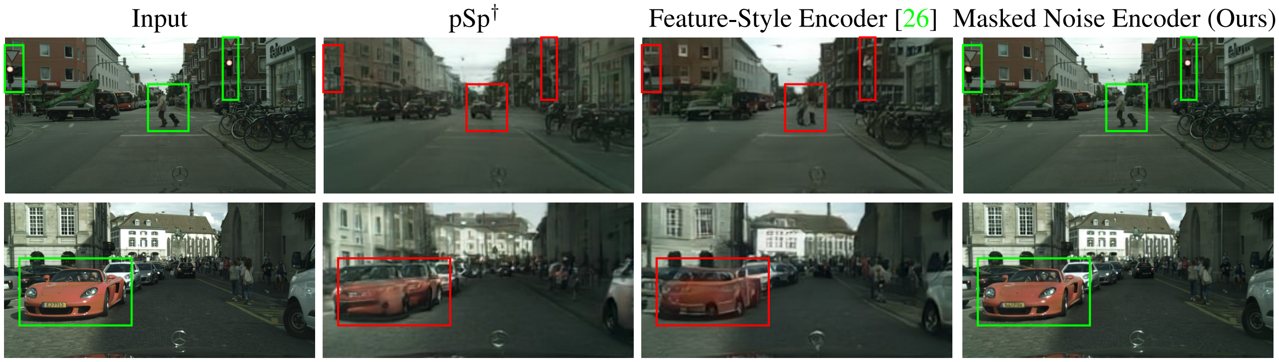


Figure 2. Qualitative results (best view in color and zoom in) of StyleGAN2 inversion methods on Cityscapes, i.e., pSp†, Feature-Style encoder [26] and our masked noise encoder. Note, pSp† is an improved version of pSp [18] introduced by us. pSp† can reconstruct the rough layout of the scene but still 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.

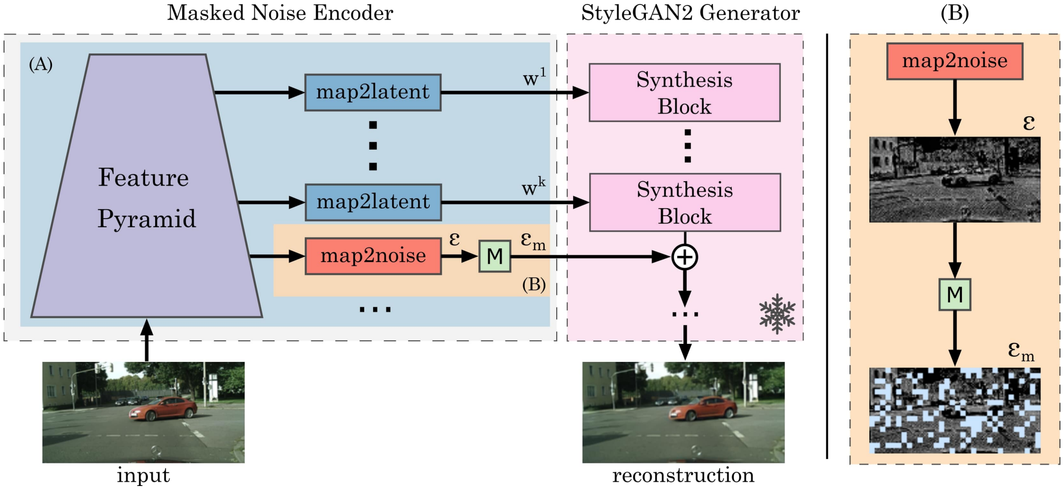


Figure 3. **Method overview.** Our encoder is built on top of the pSp encoder [18], shown in the blue area (A). In addition to mapping the input image to the extended latent space $\mathcal{W}^{+}$ of the pre-trained StyleGAN2 generator, our encoder predicts the noise map at an intermediate scale, illustrated in the orange area (B), to promote the reconstruction quality on complex scene-centric dataset, e.g., Cityscapes. $\boxed{\text{M}}$ 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.

content from style information based on a single source domain. This allows to alter the style of an image while leaving the content unchanged. Specifically, we make use of the styles extracted within the source domain and mix them up. Thus, 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 [2, 7, 18, 19, 26] are mainly tested on simple single-object-centric datasets, e.g., CelebA-HQ [13], FFHQ [15], or LSUN [28]. As shown in [1], 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.

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 explore GAN inversion for intra-source data augmentation, which can improve generalization under natural distribution shifts on semantic segmentation.

- We demonstrate ISSA can promote domain generalization performance on driving-scene semantic segmentation across different network architectures.

2. Method

2.1. Intra-Source Style Augmentation (ISSA)

ISSA employs GAN inversion to modify styles of the training samples while preserving their semantic content. In doing so, it diversifies the source training set 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 re-used for training, without requiring any further annotation effort.

StyleGAN [14–16] can synthesize scene-centric datasets like Cityscapes [5] and BDD100K [27]. However, existing GAN inversion encoders cannot provide the desired fidelity to enable ISSA to improve domain generalization of semantic segmentation via data augmentation. Loss of fine details or inauthentic reconstruction of small-scale objects would harm the model's generalization ability. Therefore, we propose a novel encoder design to invert StyleGAN2, termed *masked noise encoder* (see Fig. 3), which can faithfully reconstruct complex scenes and separately encode the style and content information.

2.2. Masked Noise Encoder

We build our encoder upon the pSp encoder [18]. It employs a feature pyramid [17] to extract multi-scale features from a given image, see Fig. 3-(A). 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.

Additive Noise Map. Due to the gap between the real and synthetic data distributions [18], we predict latent code in the extended latent space $\mathcal{W}^+$ rather than $\mathcal{W}$ of StyleGAN2. However, the latent codes $\{w^k\}$ from $\mathcal{W}^+$ alone are not expressive enough to reconstruct images with diverse semantic layouts such as Cityscapes [5] as shown in Fig. 2-(pSp†). To address this issue, our encoder additionally predicts the additive noise map ε of the StyleGAN2 at an intermediate scale, i.e., $\mathrm{map2noise}$ in Fig. 3-(B).

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 [9, 25]. In particular, we spatially divide the noise map into non-overlapping $P \times P$ patches, see $\boxed{\text{M}}$ in Fig. 3-(B). 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.

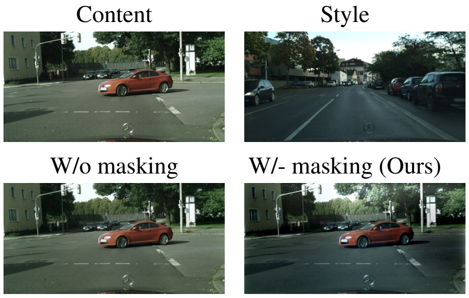


Figure 4. Style mixing effect enabled by random noise masking (best view in color). The encoder trained without masking cannot change the style of the given $\mathrm{Content}$ image. In contrast, the encoder trained with masking can modify it using the style from the given Style image.

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. Fig. 4 visualizes the style mixing effect.

3. Experiments

Datasets. We conduct extensive experiments using the following driving scene datasets: Cityscapes (CS) [5], ACDC [20]. Cityscapes is collected under good/medium weather conditions during daytime, primarily in Germany. ACDC contains four adverse weather conditions (rain, fog, snow, night) and is collected in Switzerland. The default setting is to use Cityscapes as the source training data, whereas the validation set of ACDC represents unseen target domains with different types of natural shifts, i.e., used only for testing. We consider a *single source domain* for training.

Masked Noise Encoder. Table 2 shows that our masked noise encoder considerably outperforms two strong StyleGAN2 inversion baselines pSp [18] and Feature-Style encoder [26] in all three evaluation metrics. The achieved low values of MSE, LPIPS [30] and FID [11] indicate its high-quality reconstruction. pSp† is an improved version of pSp introduced by us, which is trained with ground truth latent codes w_{gt} for better initialization. While pSp† improves over pSp in MSE and FID, it still underperforms compared to the others.

Domain Generalization. Table 1 reports the mIoU scores

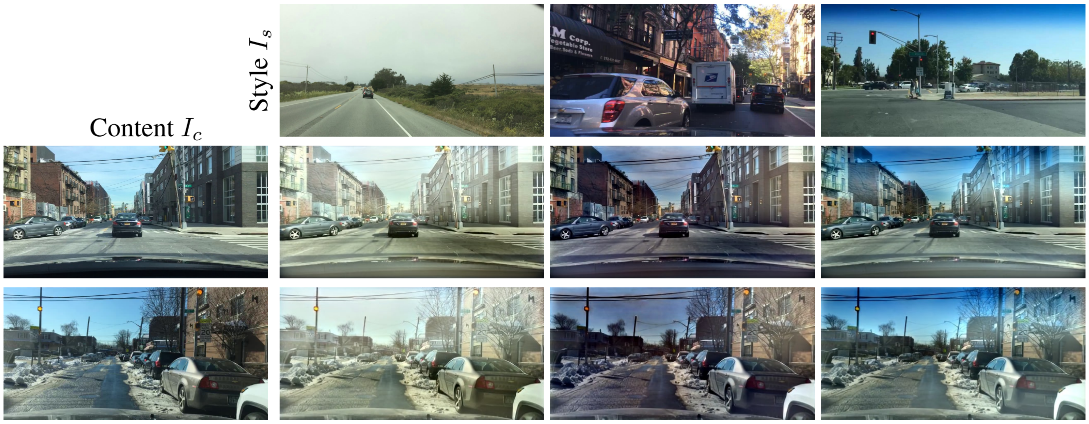


Figure 5. Visual examples of style mixing on BDD100K (best view in color) enabled by our masked noise encoder. By combining the latent codes $\{w_s^k\}$ of I_s and the noise map ε_c of I_c, the synthesized images $G(w_s^k, \varepsilon_c)$ preserve the content of I_c with a new style resembling I_s.

	HRNet [23]						SegFormer [24]					
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
CutMix [29]	**72.68**	42.48	58.63	44.50	17.07	40.67	**69.23**	49.53	61.58	47.42	27.77	46.57
Weather [10]	69.25	**50.78**	60.82	38.34	22.82	43.19	67.41	54.02	64.74	49.57	28.50	49.21
StyleMix [12]	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

Table 1. Comparison of data augmentation for improving domain generalization, i.e., from Cityscapes (train) to ACDC (unseen). 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.). 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.

Method	MSE ↓	LPIPS ↓	FID ↓
pSp [18]	0.078	0.348	130.62
pSp† [18]	0.049	0.339	14.60
Feature-Style [26]	0.025	0.220	7.14
Ours	**0.011**	**0.124**	**3.94**

Table 2. Reconstruction quality on Cityscapes at the resolution 128×256. 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.

of Cityscapes to ACDC domain generalization using two semantic segmentation models, i.e., HRNet [23] and SegFormer [24]. ISSA is compared with three representative data augmentations methods, i.e., CutMix [29], Hendrycks's weather corruptions [10], and StyleMix [12]. 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 more than 5.8% in mIoU using HRNet. StyleMix [12] also seeks to mix up styles. However, due to its poor synthetic image quality, it could even hurt the performance.

4. Conclusion

In this paper, we propose a GAN inversion based data augmentation method ISSA for learning domain generalized semantic segmentation using restricted training data from a single source domain. The key enabler for ISSA 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. We verify the effectiveness of ISSA on domain generalization across different datasets and network architectures.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *CVPR*, 2020. 2

[2] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *CVPR*, 2022. 2

[3] Simon Burton, Lydia Gauerhof, and Christian Heinzemann. Making the case for safety of machine learning in highly automated driving. In *SAFECOMP*, 2017. 1

[4] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *CVPR*, 2018. 1

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 1, 3

[6] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint*, 2017. 1

[7] Tan M Dinh, Anh Tuan Tran, Rang Nguyen, and Binh-Son Hua. Hyperinverter: Improving stylegan inversion via hypernetwork. In *CVPR*, 2022. 2

[8] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016. 1

[9] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 3

[10] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2018. 1, 4

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 3

[12] Minui Hong, Jinwoo Choi, and Gunhee Kim. StyleMix: Separating content and style for enhanced data augmentation. In *CVPR*, 2021. 4

[13] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *ICLR*, 2018. 2

[14] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *NeurIPS*, 2020. 3

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2

[16] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 1, 3

[17] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 3

[18] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *CVPR*, 2021. 2, 3, 4

[19] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *arXiv preprint*, 2021. 2

[20] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *ICCV*, 2021. 1, 3

[21] Sina Shafaei, Stefan Kugele, Mohd Hafeez Osman, and Alois Knoll. Uncertainty in machine learning: A safety perspective on autonomous driving. In *SAFECOMP*, 2018. 1

[22] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *NeurIPS*, 2020. 1

[23] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2020. 1, 4

[24] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021. 4

[25] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SimMIM: A simple framework for masked image modeling. In *CVPR*, 2022. 3

[26] Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. Feature-Style Encoder for Style-Based GAN Inversion. *arXiv preprint*, 2022. 2, 3, 4

[27] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020. 3

[28] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint*, 2015. 2

[29] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019. 1, 4

[30] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 3

[31] Yuxiao Zhang, Alexander Carballo, Hanting Yang, and Kazuya Takeda. Autonomous Driving in Adverse Weather Conditions: A Survey. *arXiv preprint*, 2021. 1