Intra-Source Style Augmentation for Improved Domain Generalization Supplementary Material

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

This supplementary material to the main paper is structured as follows:

- In Appendix S.1 Masked Noise Encoder, we include an ablation study on noise masking, and noise map resolution. More qualitative results of the encoder comparison are provided to supplement Fig. 2 in the main text. Additionally, we provide information regarding the computational complexity.
- In Appendix S.2 Domain Generalization, we present a detailed quantitative comparison with other data augmentation techniques, as well as visual results of the semantic segmentation. ISSA improves the model's generalization performance, supporting the results of Table 3 in the main paper. We also demonstrate the plug-n-play ability of ISSA.
- In Appendix S.3 Comparison with Unsupervised Domain Adaptation Methods, we show that ISSA is competitive with unsupervised domain adaptation methods, even though it does not have access to the target domain data.
- In Appendix S.4 Limitations and Future Work, we provide the discussion on the limitations and future directions of the proposed method.

S.1. Masked Noise Encoder

Ablation on noise random masking. We conduct an ablation study on the mask patch size P and masking ratio ρ, shown in Table S.1. We observe that 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 S.2, using 32×64 noise, i.e., one fourth of the image resolution, achieves better 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. S.1. Therefore, we employ the intermediate noise map at one fourth of the input resolution in all of our experiments.

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

Table S.1. Ablation on the mask patch size and masking ratio. The influence of patch size is minor on the reconstruction, while masking ratio is more important, i.e., higher masking ratio has negative impact.

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

Table S.2. Effect of noise map resolution on reconstruction quality. Experiments are done on Cityscapes, 128×256 resolution.

Additional qualitative results. In Fig. S.2 we provide more visual results of the comparison among pSp [8], pSp†, feature-style encoder [17] and our masked noise encoder. Note that, pSp† is an improved version obtained by us, which is trained with an additional discriminator and synthesized images for better initialization. It is evident that our masked noise encoder is capable of preserving more fine details and high-quality reconstruction, which is consistent with the observation in Fig. 2 in the main text.

Computational complexity. We provide more details on the time and memory usage required by using the masked noise encoder. 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 train-

Figure S.1. 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.

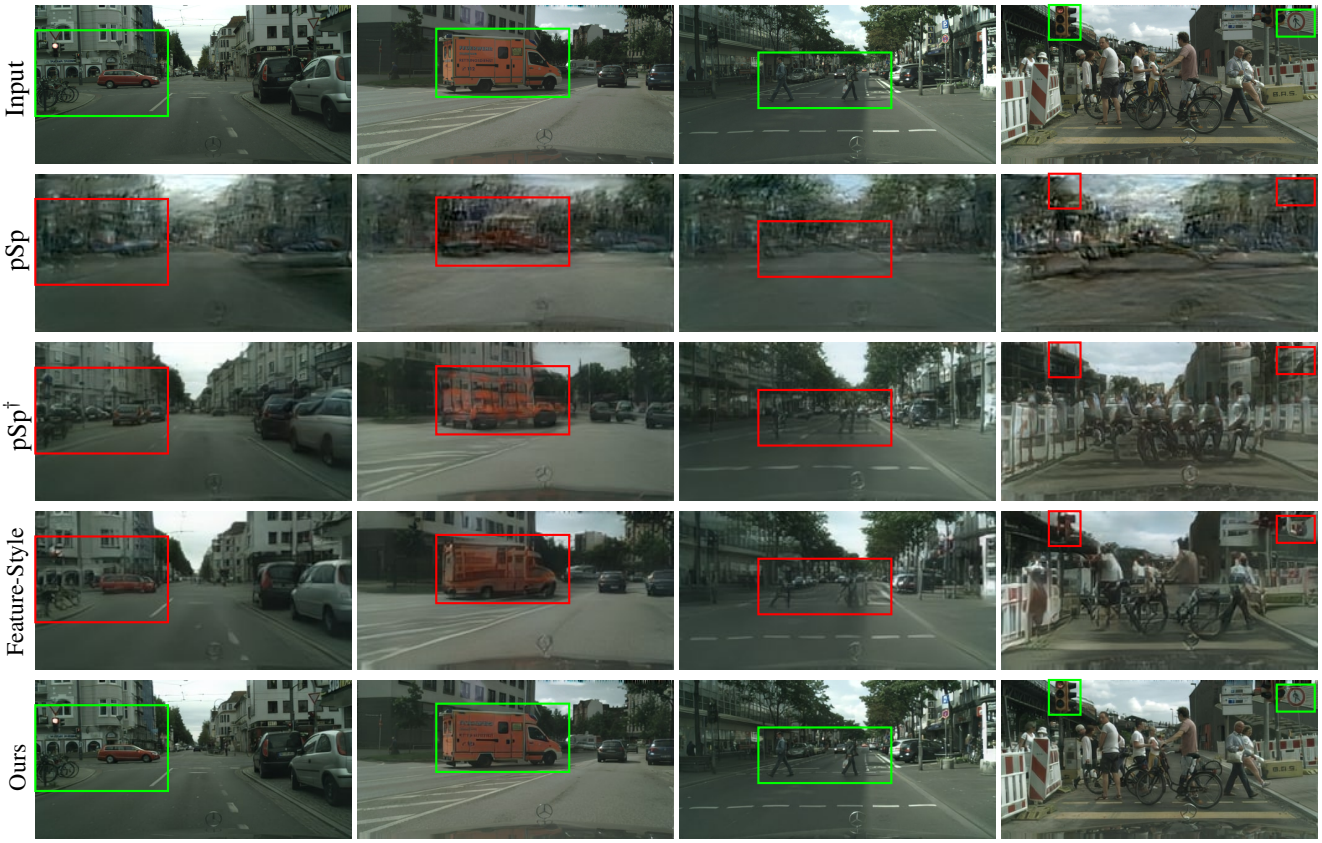

Figure S.2. Qualitative comparison between our masked noise encoder and other StyleGAN2 inversion encoders on Cityscapes (best view in color and zoom in). Note, pSp† is obtained by us, training pSp with an additional discriminator and incorporate synthesized images for better initialization. Evidently, our masked noise encoder achieves the highest fidelity and successfully reconstruct small objects such pedestrians and traffic signs. This is consistent with the observation in Fig. 2 of the main text.

ing. Nonetheless, for data augmentation, it only concerns the inference time of our encoder, which is much faster, i.e., 0.1 seconds, compared to optimization based methods such as PTI [9] that takes 55.7 seconds per image. Furthermore, stylized images by ISSA can be pre-generated and pre-stored instead of being generated on-the-fly for data augmentation, reducing the memory usage during the semantic segmentation network training.

S.2. Domain Generalization

Comparison with data augmentation methods. Table S.3 provides the full comparison on Cityscapes to ACDC domain generalization between ISSA and other data augmentation methods, e.g., CutMix [18], Hendrycks corruptions [2] and StyleMix [3]. Two semantic segmentation models HRNet [13] and SegFormer [15] are used. We report more generalization results on BDD100K and Dark Zürich in Table S.4. Supporting results in Table 3 of the main paper, ISSA has shown consistent improvements on models' generalization capability across datasets and network architectures. We also observe that, among different Hendrycks corruption types, noise and blur corruptions have larger negative impact on the performance, while weather and digital corruptions can offer little help on the generalization performance.

Method	HRNet [13] CS	Rain	Fog	Snow	Night	Avg.	SegFormer [15] 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 [18]	**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 [2]	69.25	**50.78**	60.82	38.34	22.82	43.19	67.41	54.02	64.74	49.57	28.50	49.21
Noise [2]	65.78	42.45	54.60	41.64	16.31	38.75	65.89	53.15	63.88	46.63	27.66	47.83
Digital [2]	69.13	50.13	65.71	49.22	24.81	47.47	67.57	55.53	66.46	49.92	30.33	50.56
Blur [2]	65.95	44.05	51.22	40.19	16.83	38.07	66.15	51.17	61.57	45.71	27.49	46.48
Common [2]	68.68	52.00	62.33	43.42	21.78	44.88	67.26	55.63	66.78	48.50	32.63	50.89
StyleMix [3]	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**
ISSA+CutMix	72.37	53.42	68.88	53.82	30.10	51.55	68.43	55.85	68.70	52.98	33.82	52.84
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 S.3. 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 mIoU upper bound on ACDC for the other methods. Note, it is not supposed to be an upper bound on Cityscapes. ISSA performs the best on ACDC using both HRNet and SegFormer, consistently improving the mIoU in all four scenarios of ACDC. This table complements Table 3 of the main paper with additional types of Hendrycks' corruption types, i.e., noise, digital and blur. Additionally, we combine ISSA with CutMix to diversify both styles and content of the training samples, where CutMix brings performance gain on the source domain.

Method	HRNet [13] CS	ACDC	BDD100K	Dark Zürich	SegFormer [15] CS	ACDC	BDD100K	Dark Zürich
Baseline	70.47	41.48	45.66	15.50	67.90	47.04	49.35	24.20
CutMix [18]	**72.68**	40.67	45.57	15.34	**69.23**	46.57	48.93	22.98
Weather [2]	69.25	43.19	44.53	18.71	67.41	49.21	49.84	23.44
Noise [2]	65.78	38.75	44.13	12.40	65.89	47.83	49.55	22.50
Digital [2]	69.13	47.47	47.60	22.32	67.57	50.56	51.11	25.11
Blur [2]	65.95	38.07	37.16	15.05	66.15	46.48	48.89	22.82
Common [2]	68.68	44.88	46.31	18.30	67.26	50.89	51.53	27.11
StyleMix [3]	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**
ISSA+CutMix	72.37	51.55	50.06	26.24	68.43	52.84	51.89	28.29

Table S.4. Comparison of data augmentation for improving domain generalization, i.e., from Cityscapes (train) to ACDC, BDD100K and Dark Zürich (unseen). The mean Intersection over Union (mIoU) is reported. This table supplements the results in Table 3 of the main paper. ISSA consistently outperforms the other data augmentation techniques across different datasets and network architectures. We additionally combine ISSA with CutMix to diversify both styles and content of the training samples, where CutMix brings performance gain on the source domain.

Besides, we consider BDD100K-Daytime as the source domain, ACDC and Dark Zürich as the unseen target domains. We report the quantitative results in Table S.5. As BDD100K already covers different times of day and diverse weather conditions, we only use a subset, i.e., 2526 daytime images of BDD100K for training, to allow for a more representative domain generalization evaluation. In this case, we specifically report ACDC-Night performance, since only nighttime images are not included in the training set. ISSA still outperforms the other data augmentation methods on unseen domains, being coherent with the other experimental results.

Qualitative results of ISSA. We present visual examples of our ISSA in Fig. S.4. Images in each row have the same content with random styles extracted from the source domain, i.e., Cityscapes for the 1st row and BDD100K-Daytime for the remaining rows. Besides, some qualitative semantics segmentation results on Cityscapes to ACDC generalization are demonstrated in Fig. S.5.

Plug-n-play ability. Training GAN and encoder could take considerable computational resources, therefore we instigate the plug-n-play ability of our pipeline. We observe that ISSA can still be effective even when encoder and generator are trained on a different dataset of a similar task, and re-training is not required. As shown in Table S.6, when training the segmenter on Cityscapes using ISSA, we can

Method	BDD100K	ACDC-Night	DarkZürich
Baseline [13]	52.97	23.52	23.63
CutMix [18]	**54.03**	24.37	23.99
Weather [2]	52.10	23.79	24.21
Noise [2]	49.25	19.69	19.31
Blur [2]	50.92	20.68	20.08
Digital [2]	52.10	24.17	23.24
Common [2]	51.34	23.76	23.62
StyleMix [3]	46.33	19.13	19.27
ISSA(Ours)	53.37	**25.93**	**26.55**

Table S.5. Comparison of data augmentation techniques for improving domain generalization using HRNet [13], i.e., from BDD100K-Daytime to ACDC-Night and Dark Zürich. 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.

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**

Table S.6. Comparison on Cityscapes to ACDC generalization using ISSA with generator and encoder trained on Cityscapes (CS-G-E) and BDD100K (BDD-G-E), respectively. Despite never seeing Cityscapes samples, ISSA with BDD-G-E is still highly effective.

Method	Network	Use Target	mIoU
Baseline		—	30.9
BDL [5]	DeepLabv2 [1]	✓	32.7
CRST [20]		✓	32.8
AdaptSegNet [11]		✓	33.4
SIM [14]		✓	34.6
MRNet [19]		✓	36.1
ADVENT [12]		✓	37.7
CLAN [6]		✓	39.0
FDA [16]		✓	45.7
ISSA(Ours)		✗	43.2
DAFormer [4]	DAFormer [4]	✓	55.4
ISSA(Ours)	SegFormer [15]	✗	52.5

Table S.7. Quantitative comparison on Cityscapes → ACDC with UDA methods. 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 [10].

directly use generator and encoder trained on BDD100K without fine-tuning. The effectiveness of ISSA is not compromised even though the model has never seen Cityscapes samples. Visual examples in Fig. S.3 show the plug-n-play style-mixing ability of our encoder on web-crawled images, where the model is only trained on Cityscapes.

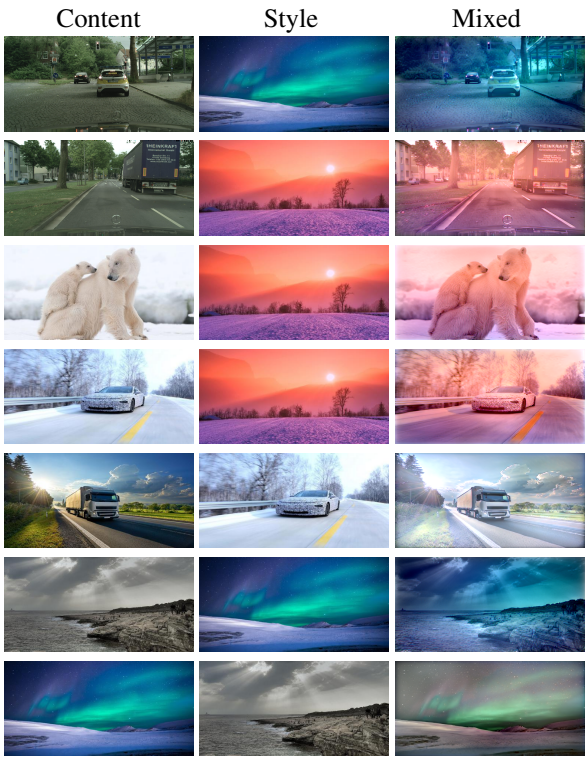

Figure S.3. Style-mixing using web-crawled images, where the generator and encoder are only trained on Cityscapes. Except for the content images of the first 2 rows, all the others are web-crawled images.

S.3. 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 S.7. Our method has presented competitive performance, even without using images from the target domain.

S.4. Limitations 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. In the future, it is challenging yet interesting to extend our work with class-aware style mixing. Also, by exploiting the pre-trained language-vision model such as CLIP [7], 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.

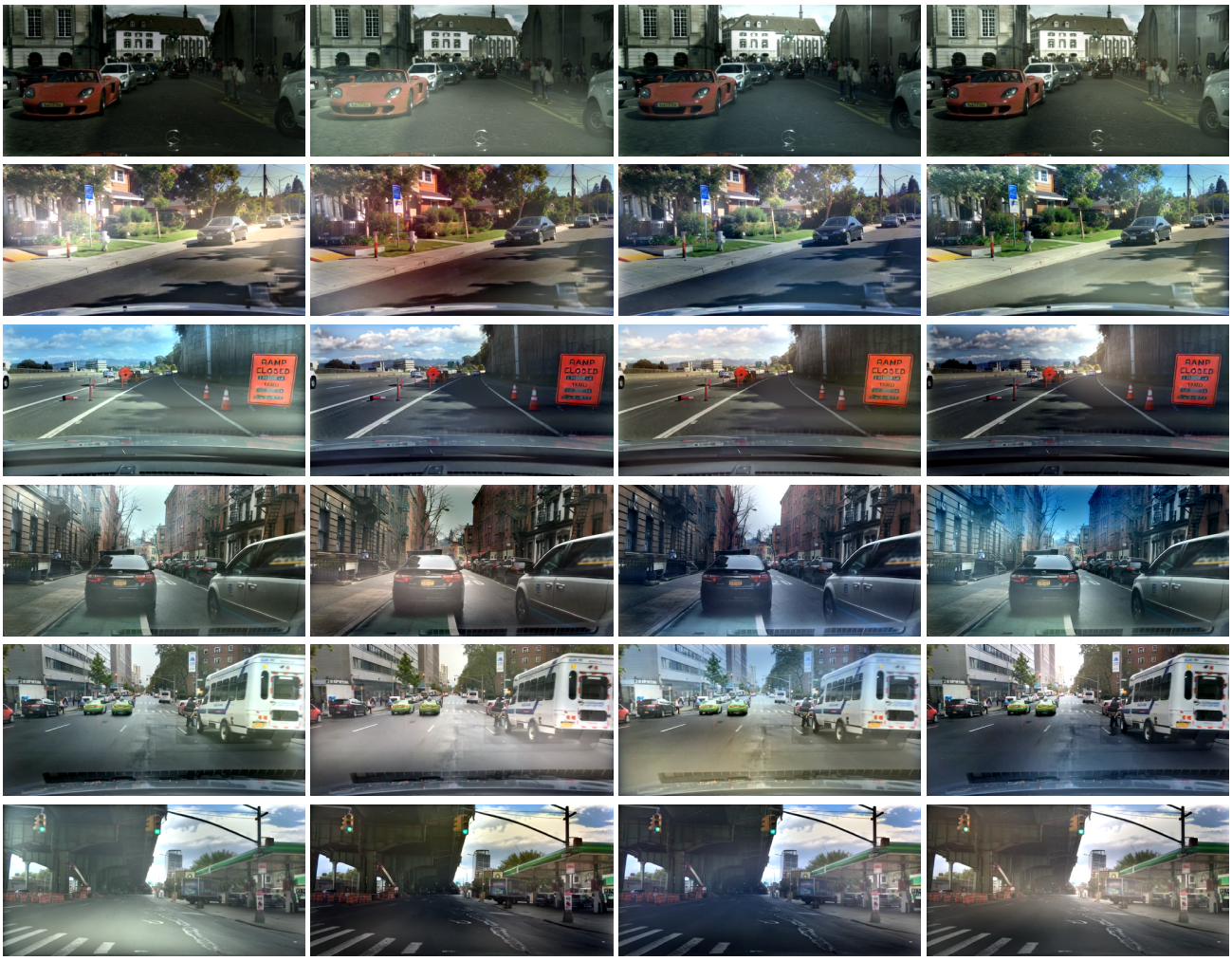

Figure S.4. Examples of augmented images by our intra-source style augmentation (ISSA). Each row presents randomly stylized samples of the same content using ISSA, where both content and styles come from the source domain only, i.e., Cityscapes for the 1st row and BDD100K-Daytime for the remaining rows.

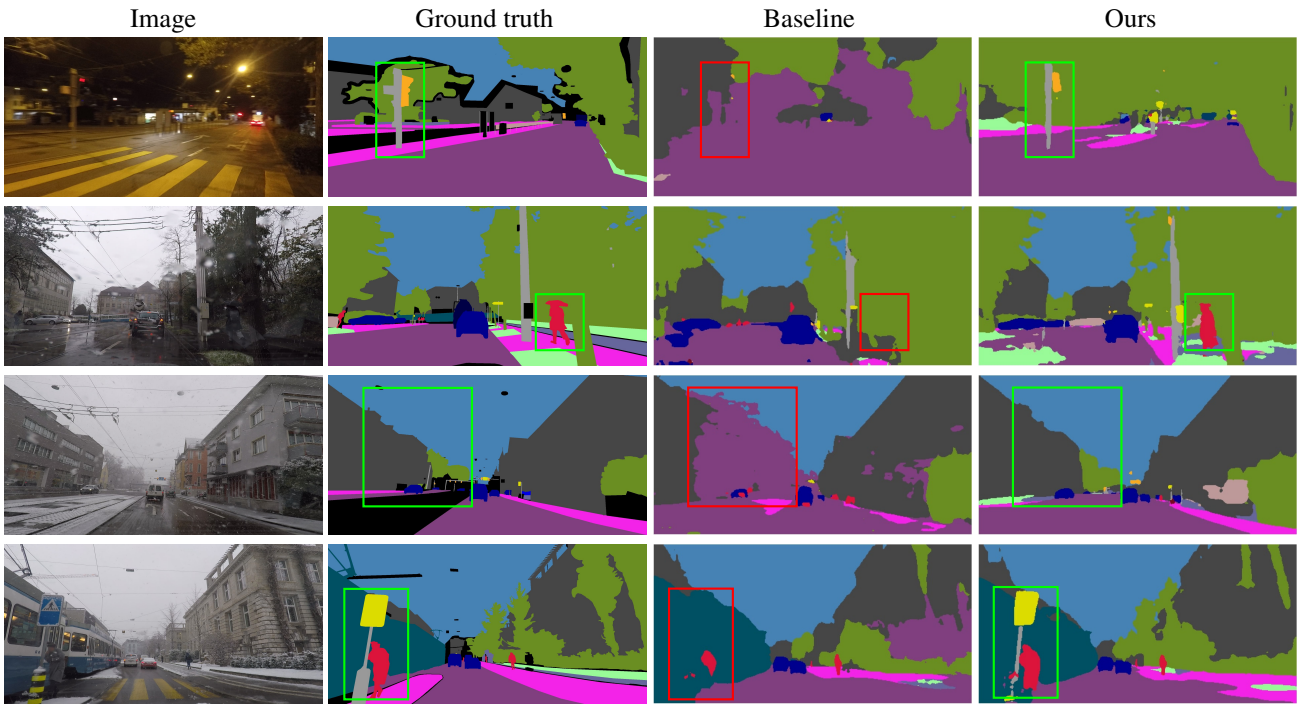

Figure S.5. Semantic segmentation results of Cityscapes → ACDC generalization using HRNet. The HRNet is trained on Cityscapes only.

References

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 2017. 4

[2] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2018. 2, 3, 4

[3] Minui Hong, Jinwoo Choi, and Gunhee Kim. StyleMix: Separating content and style for enhanced data augmentation. In *CVPR*, 2021. 2, 3, 4

[4] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *CVPR*, 2022. 4

[5] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *CVPR*, 2019. 4

[6] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *CVPR*, 2019. 4

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4

[8] 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. 1

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

[10] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *ICCV*, 2021. 4

[11] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, 2018. 4

[12] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain adaptation for structured output via discriminative patch representations. In *ICCV*, 2019. 4

[13] 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. 2, 3, 4

[14] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-mei Hwu, Thomas S Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *CVPR*, 2020. 4

[15] 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. 2, 3, 4

[16] Yanchao Yang and Stefano Soatto. FDA: Fourier domain adaptation for semantic segmentation. In *CVPR*, 2020. 4

[17] Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. Feature-Style Encoder for Style-Based GAN Inversion. *arXiv preprint*, 2022. 1

[18] 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. 2, 3, 4

[19] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *IJCV*, 2021. 4

[20] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *ICCV*, 2019. 4