In this supplementary document, we first present additional information about the diversity of the generated dataset in Sec. A. We then provide a scale analysis of the dataset in Sec. B. In Sec. C, detailed class-wise results of the proposed RCG are provided. The limitations of our approach are discussed in Sec. D. Further example predictions are showcased in Sec. E, followed by additional examples of the MRLF module in Sec. F and samples in adverse weather conditions in Sec. G.

A Diversity of the Generated Dataset

Our DGInStyle approach leverages the Style Swap and Style Prompting techniques to diversify the generated images. The diversity of training data is critical for the trained segmentation model's domain generalization. To further evaluate the diversity of the generated dataset, we employ the Frechet Inception Distance (FID) [2] and Kernel Inception Distance (KID) [1], which measure the distributional distance between two datasets. Specifically, we ablate the Style Swap and Style Prompting modules by assessing the similarity between our generated and five real-world datasets. The FID and KID scores are computed with [3] and presented in Tab. S1 and Tab. S2, respectively. A lower score indicates a smaller domain gap between the considered pair of datasets. Thus, a lower average score suggests a better coverage of the union of diverse datasets and, thus, better diversity of the generated data. The results demonstrate that both components enhance the diversity of the generated data, with the highest quality attained when both are enabled.

Table S1: Quantitative evaluation of the generated data diversity using Frechet Inception Distance (↓) between the generated data and real-world datasets. Evidently, both Style Swap and Style Prompting play important roles in bridging the gap between the generated data and each of the real datasets, a union of which represents the task-specific domain of autonomous driving.

Swap	Prompting	CS	BDD	MV	ACDC	DZ	Average
✘	✘	124.28	98.57	81.31	141.07	238.18	136.68
✔	✘	121.07	88.64	79.57	133.53	235.76	129.71
✘	✔	121.98	95.25	80.02	136.21	233.97	133.48
✔	✔	**117.05**	**88.46**	**74.81**	**128.39**	**227.69**	**127.37**

Table S2: Quantitative evaluation of the generated data diversity using Kernel Inception Distance (KID × 0.01 ↓) between the generated data and real-world datasets. The standard deviation is part of the metric computation protocol and has also been scaled down by a factor of 0.01.

Swap	Prompting	CS	BDD	MV	ACDC	DZ	Average
✘	✘	8.54 ± 0.15	5.62 ± 0.08	4.99 ± 0.14	7.95 ± 0.18	15.66 ± 0.54	8.55 ± 0.22
✔	✘	8.19 ± 0.19	4.98 ± 0.09	5.00 ± 0.15	7.40 ± 0.16	15.38 ± 0.53	8.19 ± 0.23
✘	✔	8.24 ± 0.20	5.41 ± 0.08	5.04 ± 0.13	7.50 ± 0.18	14.93 ± 0.64	8.23 ± 0.24
✔	✔	**7.86** ± 0.22	**4.90** ± 0.09	**4.98** ± 0.17	**7.16** ± 0.18	**14.36** ± 0.67	**7.85** ± 0.27

B Dataset scale analysis

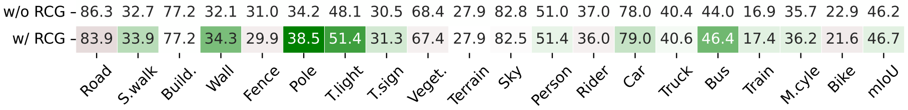
Tab. S3 studies the DG performance of DAFormer relative to the number of synthetic images. More generated images improve the mIoU up to around 6000 images, after which it reaches a plateau.

Table S3: Performance of DAFormer Using DGInStyle wrt. the unmber of generated images (mIoU ↑ in %).

$N_{\mathcal{G}}$	0	1000	2000	4000	6000	8000
Avg3	51.73	53.57	53.86	54.1	54.25	54.28
Avg5	42.18	44.95	45.86	46.22	46.47	46.39

C Class-wise results of RCG

In Fig. S1, we show the effectiveness of RCG for difficult classes, such as *pole*, *traffic light* and *bus* that have a low pixel count in the source data.



	Road	S.walk	Build.	Wall	Fence	Pole	T.light	T.sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.cyle	Bike	mIoU
w/o RCG	86.3	32.7	77.2	32.1	31.0	34.2	48.1	30.5	68.4	27.9	82.8	51.0	37.0	78.0	40.4	44.0	16.9	35.7	22.9	46.2
w/ RCG	83.9	33.9	77.2	34.3	29.9	38.5	51.4	31.3	67.4	27.9	82.5	51.4	36.0	79.0	40.6	46.4	17.4	36.2	21.6	46.7

Fig. S1: Comparison of the class-wise IoU averaged over the five datasets with and without RCG while keeping the other components of DGInStyle coupled with DAFormer. The color visualizes the difference to the first row.

D Limitations

Diffusion models exhibit a primary drawback of prolonged sampling times. As our model is based on diffusion models, it naturally inherits this slow inference property. Moreover, the proposed MRLF module operates on multiple tiles cropped from the upscaled latents, and the sampling process of all these tiles further extends the image generation duration. However, it is important to note that this extended diffusion time does not impact the inference time of the deployed segmentation networks. Furthermore, much ongoing research aims to expedite diffusion model sampling, and we believe that this issue can be alleviated through architectural advancements.

E Further Example Predictions

We present a comprehensive qualitative comparison between the predicted semantic segmentation results of HRDA trained with GTA-only data and the model trained with our DGInStyle approach. We evaluate these models on real-world datasets, including Cityscapes (*cf*. Fig. S2), BDD100K (*cf*. Fig. S3), Mapillary Vistas (*cf*. Fig. S4), ACDC (*cf*. Fig. S5), and Dark Zurich (*cf*. Fig. S6). The model trained with our DGInStyle can better segment *truck* and *bus* (as seen in Fig. S2–S5). It also exhibits a correct segmentation of *sidewalk*, effectively identifying areas that were previously overlooked by the GTA-only trained model (as seen in Fig. S2, Fig. S4). Furthermore, it enhances performance for rare classes, such as *fence* and *traffic sign* (as seen in Fig. S4). In challenging conditions, such as nighttime scenes, our DGInStyle approach significantly improves the segmentation of *sky* and *vegetation* (as seen in Fig. S5 and Fig. S6).

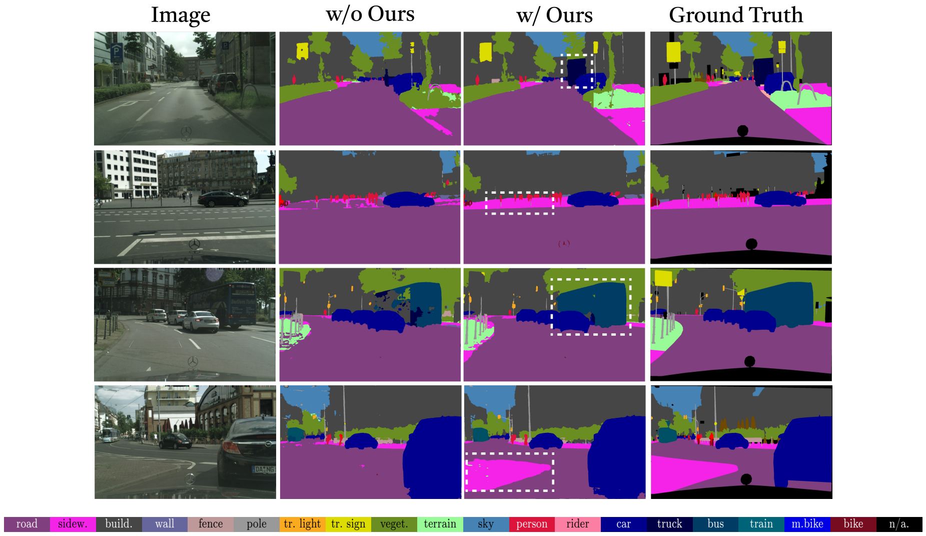


Fig. S2: Example predictions from HRDA trained with and w/o our DGInStyle on the **Cityscapes** dataset, showing improved performance on *truck* and *bus* and exhibiting a more complete segmentation of *sidewalk*.

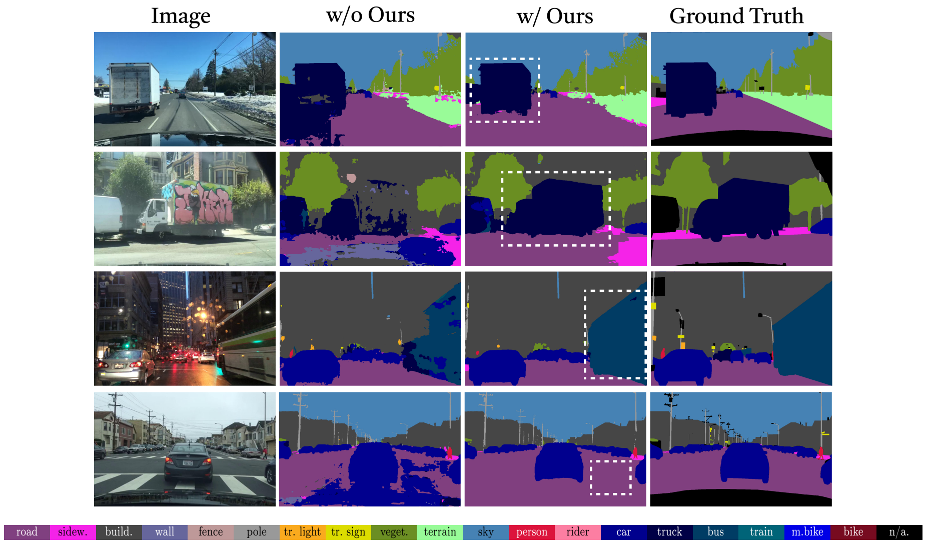


Fig. S3: Example predictions from HRDA trained with and w/o our DGInStyle on the **BDD100K** dataset, showing a better recognition of difficult classes such as *truck* and *bus*.

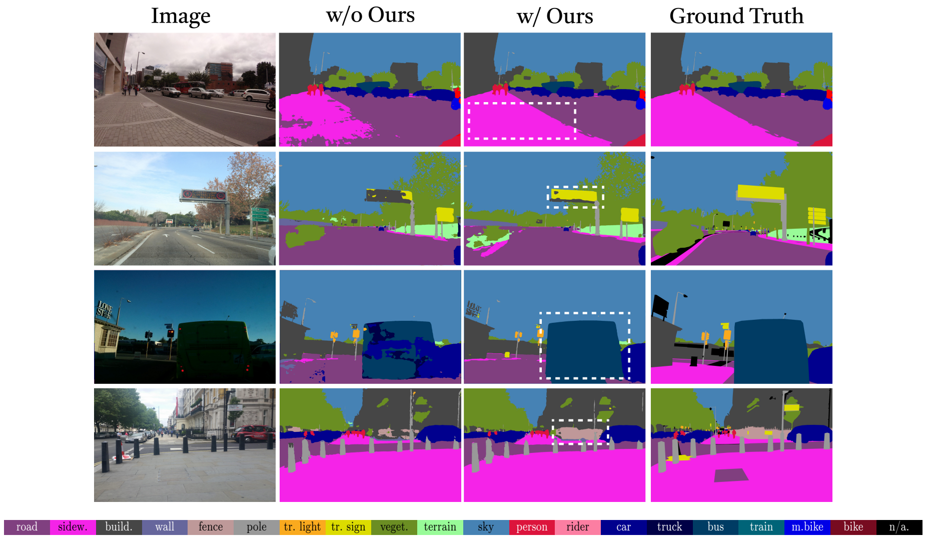


Fig. S4: Example predictions from HRDA trained with and w/o our DGInStyle on the **Mapillary Vistas** dataset, showing an improved performance of *sidewalk*, *traffic sign*, *bus* and *fence*.

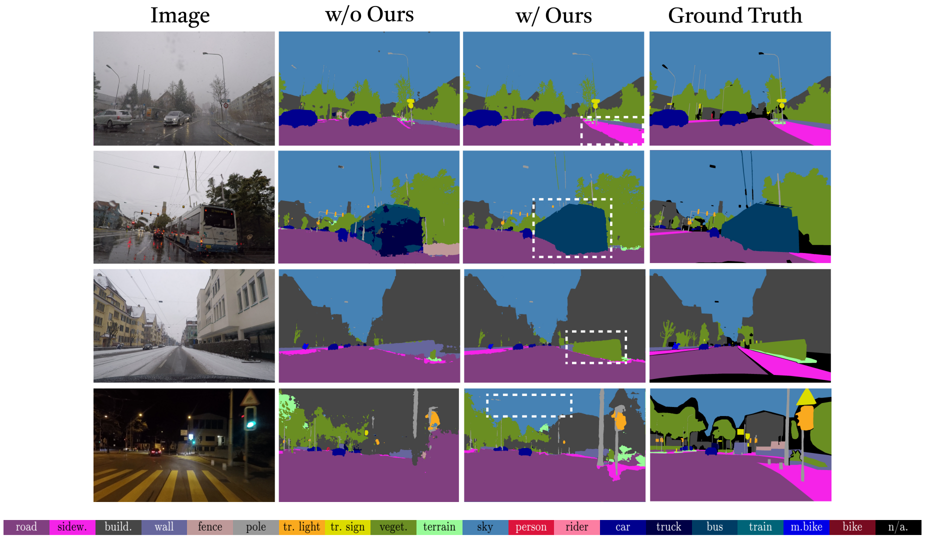


Fig. S5: Example predictions from HRDA trained with and w/o our DGInStyle on the **ACDC** dataset, demonstrating improved performance in rainy and snowy conditions for classes such as *sidewalk*, *bus*, *vegetation* and *sky*.

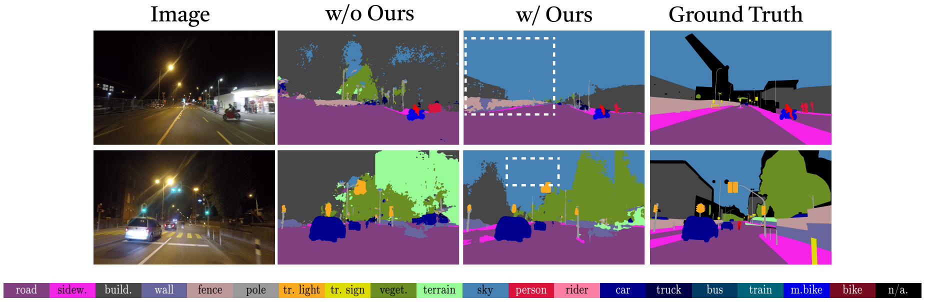


Fig. S6: Example predictions from HRDA trained with and w/o our DGInStyle on the **Dark Zurich** dataset, demonstrating superior generalization for dark scenes in the *sky* and *vegetation* classes.

F Multi-Resolution Latent Fusion Module

In Fig. S7–S9, we provide additional qualitative examples showing how the MRLF module mitigates issues of the base Stable Diffusion LDM related to the poor quality of small objects generation. For instance, in Fig. S7 (a), the motorcycle and rider are initially indistinct and poorly rendered. However, after applying the MRLF module, these elements become clearly recognizable and well-defined. Similarly, the fine-grained poles' details show a marked improvement in Fig. S8. Additionally, the quality of the person depicted in Fig. S9 also benefits significantly from the MRLF module, demonstrating its overall effectiveness in refining and improving the quality of small-scale features in generated images.

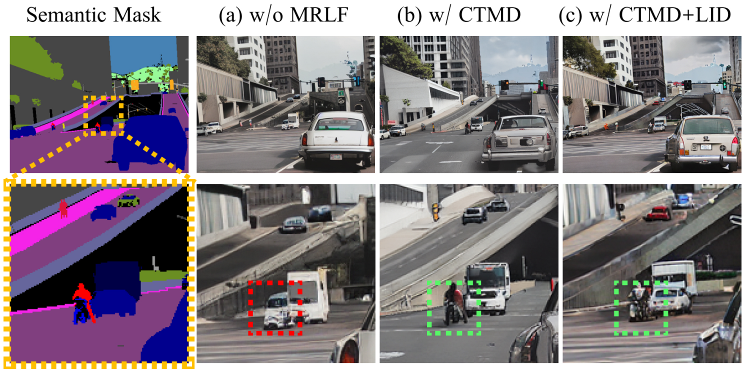


Fig. S7: Qualitative example of MRLF: improved generation for small distant objects like *rider* and *motorcycle* when zooming in on the mask crop.

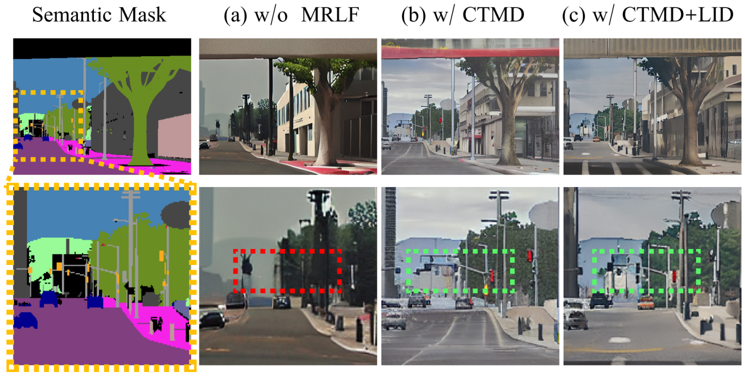


Fig. S8: Qualitative example of MRLF: improved generation for small distant objects like *pole* and *traffic light* when zooming in on the mask crop.

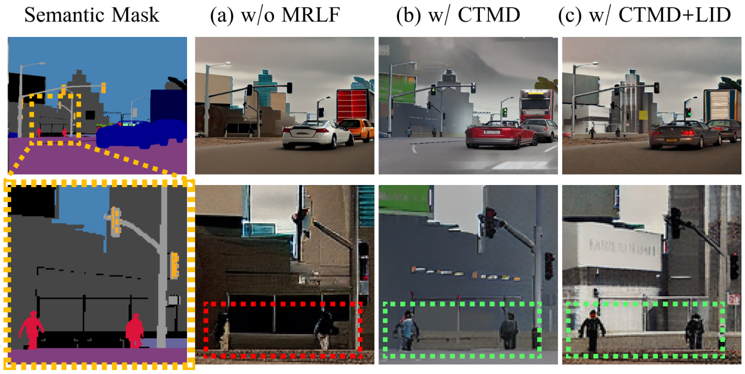


Fig. S9: Qualitative example of MRLF: improved generation for small distant objects like *person* when zooming in on the mask crop.

G Adverse Weather Samples

In Fig. S10, we show more examples of the generated content under different weather conditions given the same semantic label condition. By encompassing a wide range of weather scenarios, DGInStyle ensures that the models are well-equipped to handle real-world variations, thereby improving their applicability and reliability in diverse operational environments.

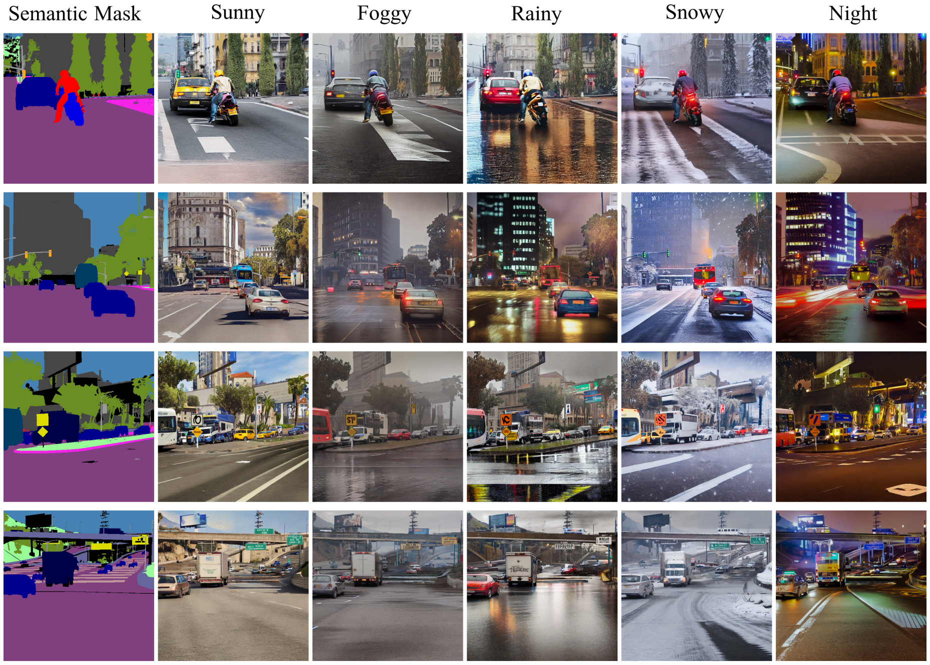


Fig. S10: Examples generated by our DGInStyle approach under varying weather conditions, all based on the same semantic label condition.

References

1. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. arXiv preprint arXiv:1801.01401 (2018)
2. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in Neural Information Processing Systems (2017)
3. Obukhov, A., Seitzer, M., Wu, P.W., Zhydenko, S., Kyl, J., Lin, E.Y.J.: High-fidelity performance metrics for generative models in pytorch (2020). `https://doi.org/10.5281/zenodo.4957738`, `https://github.com/toshas/torch-fidelity`, version: 0.3.0, DOI: 10.5281/zenodo.4957738