DGInStyle: Domain-Generalizable Semantic Segmentation with Image Diffusion Models and Stylized Semantic Control

Yuru Jia[1] Lukas Hoyer[1] Shengyu Huang[1] Tianfu Wang[1]
Luc Van Gool[1,2,3] Konrad Schindler[1] Anton Obukhov[1]

[1]ETH Zurich [2]KU Leuven [3]INSAIT Sofia

Figure 1. **Crossing domain boundaries with DGInStyle.** We propose a data-centric generative pipeline for domain generalization. It is derived from Stable Diffusion and augmented with a novel high-precision style-preserving semantic control. DGInStyle combines semantic masks (`Query`) with style prompts (e.g., `Night` or `Rain`) to generate training data for semantic segmentation networks with widely varying appearance. DGInStyle achieves state-of-the-art semantic segmentation across domains in autonomous driving.

Abstract

Large, pretrained latent diffusion models (LDMs) have demonstrated an extraordinary ability to generate creative content, specialize to user data through few-shot fine-tuning, and condition their output on other modalities, such as semantic maps. However, are they usable as large-scale data generators, e.g., to improve tasks in the perception stack, like semantic segmentation? We investigate this question in the context of autonomous driving, and answer it with a resounding "yes". We propose an efficient data generation pipeline termed DGInStyle. *First, we examine the problem of specializing a pretrained LDM to semantically-controlled generation within a narrow domain. Second, we propose a Style Swap technique to endow the rich generative prior with the learned semantic control. Third, we design a Multi-resolution Latent Fusion technique to overcome the bias of LDMs towards dominant objects. Using DGInStyle, we generate a diverse dataset of street scenes, train a domain-agnostic semantic segmentation model on it, and evaluate the model on multiple popular autonomous driving datasets. Our approach consistently increases the performance of several domain generalization methods compared to the previous state-of-the-art methods. The source code and the generated dataset are available at dginstyle.github.io.*

1. Introduction

The rise of generative image modeling has been a game changer for AI-assisted creativity. Moreover, it also paves the way for improvements beyond artistic generation, particularly in computer vision. In this paper, we investigate one such avenue and use a powerful text-to-image generative diffusion model to improve semantic segmentation.

Semantic segmentation requires large annotated datasets for supervised training. While manual annotation is costly and time-consuming [6, 40], synthetic datasets offer a cost-effective solution. However, these datasets face a *domain gap* [11], leading to poor performance when networks trained on the *source domain* are applied to the *target domain*. When the characteristics of the target domain are known, the domain gap can be addressed with Domain Adaptation techniques [11, 19]. A more challenging, arguably equally important setting is Domain Generalization (DG) [9, 21, 57], where a model is deployed in a new environment without knowing the target domain except for its general context (such as "autonomous driving").

In the DG semantic segmentation literature, the role of the *prior domain* is often overlooked or typically remains implicit [7, 21, 57]. Therefore, we take a closer look at the prior domain and study how we can utilize the rich prior that emerges in modern foundational models trained on internet-scale datasets [42] to improve domain generalization of semantic segmentation. To this end, we design **DGInStyle**,

a novel data generation pipeline with a pretrained text-to-image LDM [37] at its core, fine-tuned with source domain data and conditioned on dense label maps. Such a pipeline can automatically generate images with *characteristics of the prior domain* and *equipped with pixel-aligned label maps* (Fig. 1). The idea is that a model trained on such data will offer improved domain generalization, drawing on the prior knowledge embedded in the LDM.

This comes with two new challenges: The LDM needs to learn to produce images that match semantic masks from the labeled source domain while avoiding overfitting to its style. Additionally, the generated images need precise alignment with segmentation masks, even for very small instances. Our **Contributions** address these issues: **First** we propose a Style Swap technique inspired by modern fine-tuning and semantic style control mechanisms, to achieve the necessary level of control and diversity over the outputs. **Second**, we present a novel Multi-Resolution Latent Fusion technique, which helps us to go beyond the limited resolution of the LDM generator and achieve conditioned generation of small instances. **Lastly**, we use the resulting generative pipeline to create a diversified dataset to train semantic segmentation networks. Due to its complementary design, DGInStyle achieves major performance improvements when combined with existing DG methods. In particular, it significantly boosts the state-of-the-art domain generalization in autonomous driving.

2. Related Work

Generative Models for Dataset Generation. Diffusion Models (DMs) [3, 8, 18, 29, 43, 44] have demonstrated state-of-the-art image generation quality, primarily due to a simplified training objective. Latent diffusion models (LDMs) [37] reduce computational demands by operating in latent space, thus enabling absorption of internet-scale data [42]. A variety of diffusion models [2, 12, 14, 17, 23, 30, 54, 56] integrate additional condition signals to provide more granular control for image generation. Recent techniques have utilized DMs to create training data for downstream tasks such as image classification [1, 10, 16, 27, 41, 45], object detection [4, 48, 55], semantic segmentation [13, 26, 35, 46, 47, 51]. Paired image-mask dataset generation has been a focal point of research, with methodologies primarily falling into *grounded generation* [13, 26, 28, 46, 50], *image-to-image translation* and *Semantic guidance* [30, 51, 54]. DGInStyle falls into the last category. We use source domain masks to guide the image generation and enforce the generation fidelity using the proposed Multi-Resolution Latent Fusion technique.

Domain Generalization (DG) aims to enhance the robustness of models trained on source domains and enable them to perform well on unseen domains belonging to the same task group. To improve domain generalization in semantic segmentation, prior methods utilize transformations such as instance normalization [32] or whitening [5] to align various source domain data with a standardized feature space. Another line of research [24, 25, 33, 53, 57, 58] focuses on domain randomization, which augments the source domain with diverse styles. Recent works [13, 46] have explored the use of DMs for DG in semantic segmentation. These methods implement grounded generation by training a segmentation decoder to achieve image-mask alignment. Our approach takes a different semantic guidance route, exhibiting higher controllability and generating consistent image-label pairs that qualify as training datasets.

3. Methods

Given the labeled source domain data represented as $\mathbf{D}^S = \left\{\left(x_i^S, y_i^S\right)\right\}_{i=1}^{N_s}$, the goal is to generalize the semantic segmentation model f_θ to the unseen target domain data $\mathbf{D}^T$, by utilizing the generated labeled dataset $\mathbf{D}^G = \left\{\left(x_i^G, y_i^G\right)\right\}_{i=1}^{N_g}$. x and y stand for the images and their corresponding labels, respectively, whereas N_s and N_g are the total number of images in each dataset. $\{y_i^G\}_{i=1}^{N_g}$ is a subset of $\{y_i^S\}_{i=1}^{N_s}$ in our case, although other labels are possible.

Label Conditioned Image Generation. We use existing semantic masks and conditional text-to-image LDMs to obtain pairs of pixel-aligned images and masks. Specifically, we employ the recent work ControlNet [54] due to its efficient guidance and accessible computational requirements. During training, we convert segmentation masks into one-hot encodings, pass them as inputs to ControlNet, and supervise it with the corresponding images from the source domain. We also pass the unique class names extracted from the segmentation mask as a text prompt. Once trained, we condition the generation process on source domain masks and thus construct the new training data.

Preserving Style Prior with Style Swap. Training ControlNet from the base LDM pretrained on the prior domain leads to overfitting to the style of the domain it is fine-tuned on (as shown in Fig. 2 (c)), which limits style diversity in the generated images. To mitigate this, we develop a Style Swap technique to remove the domain-specific style in three steps, shown in Fig. 3.

(a) Source Mask (b) Source Image (c) Gen. w/o Swap (d) Gen. w/ Swap

Figure 2. **ControlNet learns the source domain style.** This effect hinders varied data generation for domain generalization. Our Style Swap mitigates the effect and preserves the style prior.

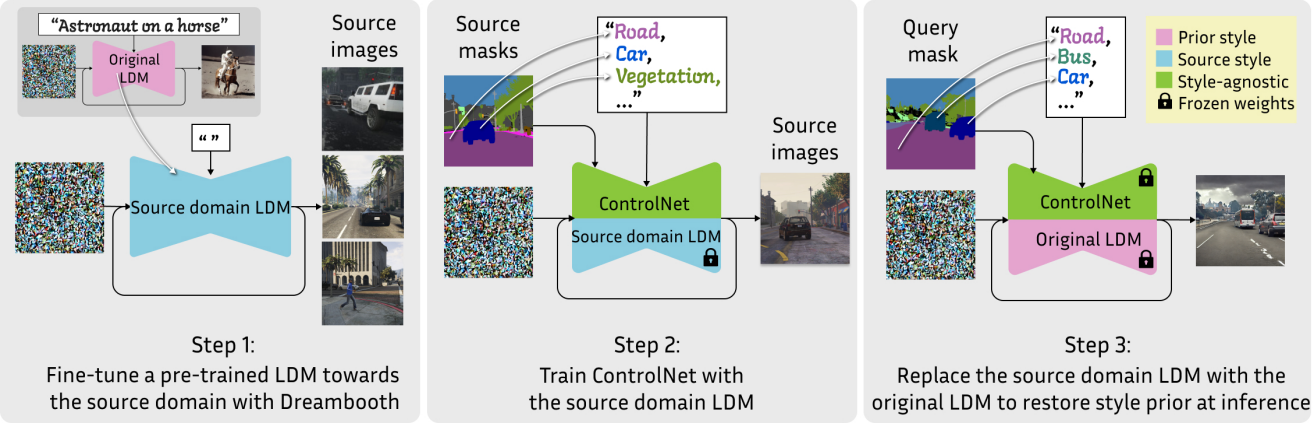

Figure 3. **Overview of our proposed Style Swap technique.** ControlNet learns segmentation-conditioned image generation on the source domain. To avoid that ControlNet also steers the generated style, it is trained on top of a source domain fine-tuned LDM. So, the source domain LDM can be replaced with the original LDM to restore its rich style prior.

As a first step of our Style Swap technique, we fine-tune the base LDM's U-Net $\mathbf{U^P}$ encapsulating the prior domain with the Dreambooth protocol [38] using source domain images, resulting in $\mathbf{U^S}$. Second, we use $\mathbf{U^S}$ as the base model instead of $\mathbf{U^P}$ to initialize ControlNet, allowing $\mathbf{U^S}$ to absorb the domain style while ControlNet focuses on the task-specific yet style-agnostic layout control. Finally, we perform inference with the trained ControlNet, except that we switch the base LDM generator to $\mathbf{U^P}$ while keeping the ControlNet trained for $\mathbf{U^S}$. This overall procedure endows the original LDM with task-specific semantic control, allowing us to generate diverse images adhering to the semantic segmentation masks (Fig. 2 (d)).

Style Prompting. We concatenate unique class names present in the semantic mask into a list and pass it to the text encoder to guide the generation. To further diversify generated data, we add randomized task-specific qualifiers to the text conditioning, including a range of adverse weather conditions (e.g., foggy, snowy, rainy, overcast, and night scenarios). An example text prompt can be: *A city street scene photo with car, road, sky, rider, bicycle, vegetation, building, in foggy weather*. This approach, especially when integrated with the Style Swap technique, allows us to produce images with precise semantic layouts and varied styles.

Multi-Resolution Latent Fusion. ControlNet struggles with small objects due to its low-resolution latent space. To improve adherence to semantic masks, we propose a two-stage Multi-Resolution Latent Fusion pipeline. The first stage involves regular ControlNet generation at original resolution, serving as a reference for the second high-resolution generation pass. In the second stage, we refine the image in an upsampled latent space, focusing on smaller details, followed by downsizing to original size. This process includes **Controlled Tiled MultiDiffusion**, where high-resolution latent codes are generated and divided into overlapping tiles for diffusion, conditioned on the corresponding semantic maps and prompts. The denoised overlapping tiles are fused subsequently to maintain consistency. Nevertheless, this can degrade large objects, which we address with **Latent Inpainting Diffusion** by retaining large objects from the first pass. This combined approach enables higher-quality generation of small objects while preserving large ones, overcoming the LDM's resolution limitations.

Rare Class Generation. We tackle imbalanced datasets by considering class distribution at both the ControlNet training and dataset generation phases. Specifically, for each class c with frequency f_c in the source domain, its sampling probability is $P(c) = e^{(1-f_c)/T} / \sum_{c'=1}^{C} e^{(1-f_{c'})/T}$, where C is the total number of classes, and T controls the smoothness of the class distribution. This approach helps the model better recognize these challenging classes and also plays a role in creating a more balanced dataset.

4. Experiments

Datasets. Following [19, 21], we use GTA [36] as the synthetic source dataset and employ five real-world datasets [6, 31, 39, 40, 52] for evaluation.

Comparison with State-of-the-Art DG. In Tab. 1, we benchmark several DG methods trained using either the GTA dataset alone or augmented with our DGInStyle and subsequently evaluated across five real-world datasets to measure their generalization capability. Specifically, we integrate DGInStyle into IBN-Net [32], RobustNet [5], ColorAug (random brightness, contrast, saturation, and hue), DAFormer [19, 21], and HRDA [20, 21] covering ResNet-101 [15] and MiT-B5 [49] network architectures. The results in Tab. 1 indicate that DGInStyle significantly enhances the DG performance across various DG methods and architectures, with improvements ranging from +2.5 to +7.2

DG Method	DGInStyle	CS [6]	BDD [52]	MV [31]	Avg3	ACDC [40]	DZ [39]	Avg5	ΔAvg5
ResNet-101 [15]									
IBN-Net [32]	✘	37.37	34.21	36.81	36.13	25.85	6.12	28.07	↑ 5.1
	✔	**40.80**	**38.98**	**43.20**	**40.99**	**31.68**	**11.19**	**33.17**	
RobustNet [5]	✘	37.20	33.36	35.57	35.38	24.80	5.49	27.28	↑ 6.8
	✔	**41.03**	**39.62**	**44.85**	**41.83**	**32.30**	**12.73**	**34.11**	
DRPC [53]	✘	42.53	38.72	38.05	39.77	–	–	–	
FSDR [22]	✘	44.80	41.20	43.40	43.13	24.77	**9.66**	32.77	
GTR [33]	✘	43.70	39.60	39.10	40.80	–	–	–	
SAN-SAW [34]	✘	45.33	41.18	40.77	42.23	–	–	–	
AdvStyle [58]	✘	44.51	39.27	43.48	42.42	–	–	–	
SHADE [57]	✘	**46.66**	**43.66**	**45.50**	**45.27**	**29.06**	8.01	**34.58**	
HRDA [20, 21]	✘	39.63	38.69	42.21	40.18	26.08	7.80	30.88	↑ 7.2
	✔	**46.89**	**42.81**	**50.19**	**46.63**	**34.19**	**16.16**	**38.05**	
MiT-B5 [49]									
Color-Aug	✘	46.64	45.45	49.04	47.04	36.10	16.37	38.72	↑ 3.3
	✔	**50.76**	**47.21**	**52.33**	**50.10**	**38.92**	**20.94**	**42.03**	
DAFormer [19, 21]	✘	52.65	47.89	54.66	51.73	38.25	17.45	42.18	↑ 4.3
	✔	**55.31**	**50.82**	**56.62**	**54.25**	**44.04**	**25.58**	**46.47**	
HRDA [20, 21]	✘	57.41	49.11	61.16	55.90	44.04	20.97	46.54	↑ 2.5
	✔	**58.63**	**52.25**	**62.47**	**57.78**	**46.07**	**25.53**	**48.99**	

Table 1. **DG with GTA source domain and ResNet-101/MiT-B5 backbone.** Comparison of DG methods for semantic segmentation in autonomous driving scenes w/ and w/o integrating our generated dataset (mIoU ↑ in %) .

DG Method	DGInStyle	BDD [52]	MV [31]	ACDC [40]	DZ [39]	Average	ΔAverage
Color-Aug	✘	53.33	**60.06**	52.38	23.00	47.19	↑ 2.1
	✔	**55.18**	59.95	**55.19**	**26.83**	**49.29**	
DAFormer [19, 21]	✘	54.19	61.67	55.15	28.28	49.82	↑ 1.5
	✔	**56.26**	**62.67**	**57.74**	**28.55**	**51.31**	
HRDA [20, 21]	✘	58.49	**68.32**	59.70	31.07	54.40	↑ 0.7
	✔	**58.84**	67.99	**61.00**	**32.60**	**55.11**	

Table 2. **DG with Cityscapes source domain and MiT-B5 [49] backbone.** Cityscapes to other datasets domain generalization w/ and w/o integrating our generated dataset (mIoU ↑ in %).

mIoU on average across datasets. These results confirms the efficacy of our method in generating diverse, style-varied and accurate image-label pairs.

To broaden the scope of our evaluation, we set an experiment with Cityscapes [6] as a source domain, generalizing to other real-world domains in Tab. 2. The results in this real-to-real adaptation scenario again confirm that our generated dataset consistently boosts the performance of semantic segmentation models across all configurations.

Ablation Studies. We conduct ablation studies to evaluate each component of our method using the DAFormer framework, with results in Tab. 3. Adding Multi-Resolution Latent Fusion (MRLF) improved small object generation, boosting average performance by +0.96 across five datasets. The Style Swap technique, key to style diversification, further increased performance by +1.57, showcasing the value of leveraging prior domain knowledge for diverse sample generation. Contributions from Style Prompts and Rare Class Generation (RCG) also led to performance gains.

Modules				mIoU↑	
MRLF	Swap	Prompts	RCG	Avg3	Avg5
✘	✘	✘	✘	51.46	43.31
✔	✘	✘	✘	52.84	44.27
✔	✔	✘	✘	53.85	45.84
✔	✔	✔	✘	53.95	46.16
✔	✔	✔	✔	**54.25**	**46.67**

Table 3. **Ablation studies** on different components for our data generation pipeline. All models use DAFormer [19].

MRLF Modules		mIoU↑	
CTMD	LID	Avg3	Avg5
✘	✘	53.07	45.19
✔	✘	54.05	45.60
✔	✔	**54.25**	**46.67**

Table 4. **MRLF Ablation** with Controlled Tiled MultiDiffusion (CTMD) and Latent Inpainting Diffusion (LID).

Semantic Mask (a) w/o MRLF (b) w/ CTMD (c) w/ CTMD+LID

Figure 4. **Qualitative examples of MRLF.** (a) When zooming in on the mask crop, the initial generation fails to create recognizable content for small instances like cars and traffic poles. (b) This is addressed by conducting Controlled Tiled MultiDiffusion, which enhances the generation quality of fine details. However, it can lead to artifacts of large objects. (c) When adding Latent Inpainting Diffusion, the generated image not only improves the local details but also reduces artifacts in large objects.

To gain further insights on MRLF, we ablate its two passes while incorporating all other components during dataset generation. As shown in Tab. 4, both the Controlled Tiled MultiDiffusion (CTMD) and the Latent Inpainting Diffusion (LID) contribute to the overall performance of our method. This is also exemplified in Fig. 4, where it becomes evident that the MRLF module not only refines local details but also minimizes artifacts in larger objects.

5. Conclusion

We have explored the potential of generative data augmentation using pretrained LDMs in the challenging context of domain generalization for semantic segmentation. We propose DGInStyle, a novel and efficient data generation pipeline that crafts diverse task-specific images by sampling the rich prior of a pretrained LDM, while ensuring precise adherence of the generation to semantic layout condition. DGInStyle consistently improves the performance of several domain generalization methods across CNN and Transformer architectures, notably enhancing the state of the art. We hope that it can lay the foundation for future work on how to best utilize generative models to advance domain generalization of dense scene understanding.

References

[1] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves ImageNet classification. *arXiv:2304.08466*, 2023. 2

[2] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. *arXiv:2302.07121*, 2023. 2

[3] Shengqu Cai, Eric Ryan Chan, Songyou Peng, Mohamad Shahbazi, Anton Obukhov, Luc Van Gool, and Gordon Wetzstein. Diffdreamer: Towards consistent unsupervised single-view scene extrapolation with conditional diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2139–2150, 2023. 2

[4] Kai Chen, Enze Xie, Zhe Chen, Yibo Wang, Lanqing Hong, Zhenguo Li, and Dit-Yan Yeung. GeoDiffusion: Text-prompted geometric control for object detection data generation. *arXiv:2306.04607*, 2023. 2

[5] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2, 3, 4

[6] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1, 3, 4

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, 2009. 1

[8] Prafulla Dhariwal and Alex Nichol. Diffusion models beat GANs on image synthesis. *arXiv:2105.05233*, 2021. 2

[9] Jian Ding, Nan Xue, Gui-Song Xia, Bernt Schiele, and Dengxin Dai. HGFormer: Hierarchical grouping transformer for domain generalized semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[10] Lisa Dunlap, Alyssa Umino, Han Zhang, Jiezhi Yang, Joseph E. Gonzalez, and Trevor Darrell. Diversify your vision datasets with automatic diffusion-based augmentation. *arXiv:2305.16289*, 2023. 2

[11] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, 2015. 1

[12] Vidit Goel, Elia Peruzzo, Yifan Jiang, Dejia Xu, Nicu Sebe, Trevor Darrell, Zhangyang Wang, and Humphrey Shi. PAIR-Diffusion: Object-level image editing with structure-and-appearance paired diffusion models. *arXiv:2303.17546*, 2023. 2

[13] Rui Gong, Martin Danelljan, Han Sun, Julio Delgado Mangas, and Luc Van Gool. Prompting diffusion representations for cross-domain semantic segmentation. *arXiv:2307.02138*, 2023. 2

[14] Cusuh Ham, James Hays, Jingwan Lu, Krishna Kumar Singh, Zhifei Zhang, and Tobias Hinz. Modulating pretrained diffusion models for multimodal image synthesis. *arXiv:2302.12764*, 2023. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3, 4

[16] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? *arXiv:2210.07574*, 2023. 2

[17] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv:2207.12598*, 2022. 2

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020. 2

[19] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 3, 4

[20] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. HRDA: Context-aware high-resolution domain-adaptive semantic segmentation. *arXiv:2204.13132*, 2022. 3, 4

[21] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Domain adaptive and generalizable network architectures and training strategies for semantic image segmentation. *arXiv:2304.13615*, 2023. 1, 3, 4

[22] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 4

[23] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv:2302.09778*, 2023. 2

[24] Wei Huang, Chang Chen, Yong Li, Jiacheng Li, Cheng Li, Fenglong Song, Youliang Yan, and Zhiwei Xiong. Style projected clustering for domain generalized semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[25] Sunghwan Kim, Dae-hwan Kim, and Hoseong Kim. Texture learning domain randomization for domain generalized segmentation. *arXiv preprint arXiv:2303.11546*, 2023. 2

[26] Neehar Kondapaneni, Markus Marks, Manuel Knott, Rogério Guimarães, and Pietro Perona. Text-image alignment for diffusion-based perception. *arXiv:2310.00031*, 2023. 2

[27] Zheng Li, Yuxuan Li, Penghai Zhao, Renjie Song, Xiang Li, and Jian Yang. Is synthetic data from diffusion models ready for knowledge distillation? *arXiv:2305.12954*, 2023. 2

[28] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Guiding text-to-image diffusion model towards grounded generation. *arXiv preprint arXiv:2301.05221*, 2023. 2

[29] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2

[30] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv:2302.08453*, 2023. 2

[31] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *IEEE International Conference on Computer Vision*, 2017. 3, 4

[32] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via IBN-Net. In *European Conference on Computer Vision*, 2018. 2, 3, 4

[33] Duo Peng, Yinjie Lei, Lingqiao Liu, Pingping Zhang, and Jun Liu. Global and local texture randomization for synthetic-to-real semantic segmentation. *IEEE Transactions on Image Processing*, 30:6594–6608, 2021. 2, 4

[34] Duo Peng, Yinjie Lei, Munawar Hayat, Yulan Guo, and Wen Li. Semantic-aware domain generalized segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 4

[35] Duo Peng, Ping Hu, Qiuhong Ke, and Jun Liu. Diffusion-based image translation with label guidance for domain adaptive semantic segmentation. In *IEEE/CVF International Conference on Computer Vision*, 2023. 2

[36] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European Conference on Computer Vision*, 2016. 3

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[38] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv:2208.12242*, 2023. 3

[39] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *IEEE International Conference on Computer Vision*, 2019. 3, 4

[40] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *IEEE/CVF International Conference on Computer Vision*, 2021. 1, 3, 4

[41] Mert Bulent Sariyildiz, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. *arXiv:2212.08420*, 2023. 2

[42] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 2022. 1, 2

[43] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, 2022. 2

[44] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv:2011.13456*, 2021. 2

[45] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. *arXiv:2302.07944*, 2023. 2

[46] Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. DatasetDM: Synthesizing data with perception annotations using diffusion models. *arXiv:2308.06160*, 2023. 2

[47] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. *arXiv preprint arXiv:2303.11681*, 2023. 2

[48] Zhenyu Wu, Lin Wang, Wei Wang, Tengfei Shi, Chenglizhao Chen, Aimin Hao, and Shuo Li. Synthetic data supervised salient object detection. In *ACM International Conference on Multimedia*, pages 5557–5565, 2022. 2

[49] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 2021. 3, 4

[50] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. *arXiv:2303.04803*, 2023. 2

[51] Lihe Yang, Xiaogang Xu, Bingyi Kang, Yinghuan Shi, and Hengshuang Zhao. Freemask: Synthetic images with dense annotations make stronger segmentation models. *arXiv preprint arXiv:2310.15160*, 2023. 2

[52] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 3, 4

[53] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *IEEE/CVF International Conference on Computer Vision*, 2019. 2, 4

[54] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 2

[55] Manlin Zhang, Jie Wu, Yuxi Ren, Ming Li, Jie Qin, Xuefeng Xiao, Wei Liu, Rui Wang, Min Zheng, and Andy J. Ma. DiffusionEngine: Diffusion model is scalable data engine for object detection. *arXiv:2309.03893*, 2023. 2

[56] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K. Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *arXiv:2305.16322*, 2023. 2

[57] Yuyang Zhao, Zhun Zhong, Na Zhao, Nicu Sebe, and Gim Hee Lee. Style-hallucinated dual consistency learning for domain generalized semantic segmentation. In *European Conference on Computer Vision*, pages 535–552. Springer, 2022. 1, 2, 4

[58] Zhun Zhong, Yuyang Zhao, Gim Hee Lee, and Nicu Sebe. Adversarial style augmentation for domain generalized urban-scene segmentation. *Advances in Neural Information Processing Systems*, 2022. 2, 4