Improving End-To-End Autonomous Driving with Synthetic Data from Latent Diffusion Models

Harsh Goel
University of Texas Austin
harshg99@utexas.edu

Sai Shankar Narasimhan
University of Texas Austin
nsaishankar@utexas.edu

Sandeep Chinchali
University of Texas Austin
sandeepc@utexas.edu

Abstract

The autonomous driving field has seen notable progress in segmentation and planning model performance, driven by extensive datasets and innovative architectures. Yet, these models often struggle when encountering rare subgroups, such as rainy conditions. Obtaining the necessary large and diverse datasets to improve generalization in these subgroups is further hindered by the high cost and effort of manual annotation. To tackle this, we introduce SynDiff-AD a unique data generation pipeline designed to synthesize realistic images for under-represented subgroups. Our system utilizes latent diffusion models (LDMs) with meticulous text prompts to generate images from existing dataset annotations, faithfully preserving their semantic structure. This crucially eliminates the need for manual labeling. By augmenting the original dataset with images generated by our system, we demonstrate the improved performance of advanced segmentation models like Mask2Former and SegFormer, particularly in terms of the Intersection over Union (IoU) metric within under-represented subgroups. We also observe enhanced driving capabilities in end-to-end autonomous planning models like AIM-2D and AIM-BEV across diverse conditions. Our thorough analysis highlights that our method also contributes to overall model improvement.

1. Introduction

The field of Autonomous Driving (AD) has made significant strides in recent years, fueled by the massive datasets collected by vehicle fleets. This data explosion offers invaluable training resources. However, a key issue emerges: the collected data often exhibits a bias towards favorable weather conditions, such as sunny and clear skies. Datasets like the Waymo Open Dataset [34] and BDD100K [41] illustrate this, with a substantial portion of their samples falling into this category. This imbalance impacts model performance on segmentation [17] and planning tasks [1, 2] when encountering less common scenarios like rain or clouds. Furthermore, the associated manual labeling costs are substantial for collected extra data.

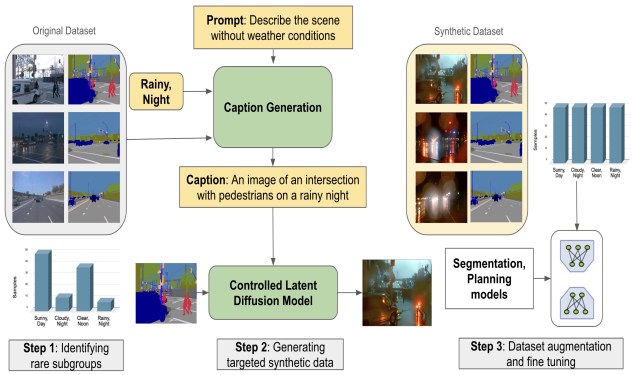

Figure 1. **Our proposed synthetic data generation method SynDiff-AD.** We begin by identifying rare subgroups in the dataset. We propose a sophisticated caption generation scheme to direct image generation with ControlNet towards rare subgroups.

To address these challenges, we propose SynDiff-AD, a novel pipeline to generate realistic data for under-represented subgroups, to augment the existing AD datasets. SynDiff-AD leverages the vast improvements in the field of text-to-image controlled generation such as [28, 30] to generate variants of the existing dataset samples while preserving their semantic structure. Our approach outlined in Fig. 1, eliminates the need for manual labeling in semantic segmentation and end-to-end autonomous driving tasks. First, we pinpoint under-represented subgroups within a dataset. We then employ ControlNet [42], a latent diffusion model (LDM), to transform samples from over-represented subgroups into the desired under-represented conditions while safeguarding their semantic content. The generated images are used to augment the existing AD dataset, allowing fine-tuning of segmentation and end-to-end driving models. Crucially, our approach demands no human labeling effort.

Our key contributions are summarised as follows:

1. We present SynDiff-AD, a conditional LDM-based method inspired by ControlNet that synthesizes sam-

ples for under-represented dataset subgroups. Our system modifies images from over-represented subgroups, meticulously preserving semantic structure.

2. Drawing inspiration from textual inversion concepts, we design a novel prompting scheme for data augmentation. We leverage LLaVA, a vision-language captioning model, to describe images within over-represented subgroups. These descriptions are carefully modified to reflect under-represented scenarios, guiding the generation process in our system.
3. Our extensive experiments show that fine-tuning state-of-the-art semantic segmentation and autonomous driving models with the augmented dataset yields improved average performance across all subgroups.

2. Related Work

2.1. Semantic Segmentation and End to End (E2E) Autonomous Driving

Prior research in semantic segmentation and E2E autonomous driving has primarily emphasized architectural modifications. State-of-the-art semantic segmentation models [5, 6, 23, 38] now heavily utilize the Transformer architecture [36]. In end-to-end autonomous driving, the focus has been on designing model architectures to predict future waypoints for vehicle tracking, often drawing on multimodal data like camera images and LiDAR depth information [4, 8, 9, 25, 32]. In contrast, our paper explores a different direction: improving the performance of state-of-the-art semantic segmentation and E2E autonomous driving models through the use of synthetic data augmentation.

2.2. Generative Models to improve Synthetic Data for Perception and Planning tasks

Text-to-image generative models built on Diffusion Probabilistic Models (DPMs), such as Stable Diffusion [30] and Glide [26], excel at generating high-quality synthetic data from text. However, they struggle with precise modifications of objects within an image. Solutions have emerged, including Pivotal Tuning [29] for targeted adjustments, Textual Inversion [12, 31] for adapting images to new concepts, and ControlNet [42] for conditioning generation on elements like semantic maps. FreestyleNet [39] further refines control through attention mapping, while Edit-Anything [13] combines tools like ControlNet, Segment Anything (SAM) [18], and BLIP2 [20] for versatile editing. These methods have been adapted to improve perception models that is complementary to GAN-generated synthetic data to improve perception models [11, 43]. Diffusion models have expanded synthetic data applications to object classification [3, 15], detection [21], and segmentation [16, 19, 35, 37, 40, 44]. Methods like DatasetDM [37] generate both images and annotations, and DGInStyle [16] focuses on style transfer. However, none specifically address segmentation performance on under-represented subgroups. For autonomous planning, synthetic data traditionally comes from expensive 3D game engines (Unity, UE4), facilitating procedural generation and expert data collection [33]. In contrast, our approach leverages existing semantic data for image generation, eliminating the need for additional human expert input.

3. Method

We outline our proposed approach, SynDiff-AD, in this section. The problem statement is described in Section 3.1. Our method for generating synthetic data conditioned on semantic masks for under-represented subgroups is detailed in Section 3.2. We enhance the quality of generated data using ControlNet fine-tuning with text captions and style prompting (Section 3.3). Finally, Section 3.4 outlines the procedure to synthesize data targeted to under-represented subgroups for semantic segmentation and autonomous driving tasks.

3.1. Preliminaries

Our goal is to enhance the performance $\mathcal{P}_f(z;\theta) = \mathbb{E}_{p(\mathbf{x},\mathbf{y}|z)} L(f(\mathbf{x};\theta),\mathbf{y})$ of a model f parameterized by θ across each subgroup z, where $p(\mathbf{x},\mathbf{y}|z)$ represents the data distribution corresponding to subgroup z. In this paper, $f(\mathbf{x};\theta)$ refers to a parameterized segmentation or autonomous driving model that produces an output $\mathbf{y}$ for input image(s) $\mathbf{x}$. For segmentation, $\mathbf{y}$ is a segmentation map and for E2E driving, $\mathbf{y}$ denotes the position offset. Additionally, $L(f(\mathbf{x};\theta),\mathbf{y})$ is a metric which is the Mean Intersection over Union (mIoU) and normalized driving score [8, 9] for segmentation and E2E driving respectively.

We assume a predefined set of subgroups $\mathcal{Z}$ which comprise an *operational design domain* [10]. Each subgroup within $\mathcal{Z}$ is a semantic attribute, for instance weather or lighting conditions. The domain is compositional [24], i.e. $\mathcal{Z} = \mathcal{Z}_0 \times \mathcal{Z}_1 \times .... \times \mathcal{Z}_{n_Z-1}$, where each $\mathcal{Z}_i$ represents a distinct semantic dimension. Here, n_Z denotes the total number of semantic dimensions. We use CLIP [27] to partition existing test datasets based on these semantic attributes, as they lack explicit subgroup labels. Evaluation uses corresponding training images if a subgroup is underrepresented in the test set.

3.2. Label Conditioned Image Generation

SynDiff-AD leverages semantic masks with text-to-image generative models to synthesize pixel-aligned image-mask pairs for under-represented subgroups. We utilize ControlNet [42], which extends Stable Diffusion [30] by providing additional control, like semantic masks, during the Latent Diffusion Model's (LDM) denoising process. This is

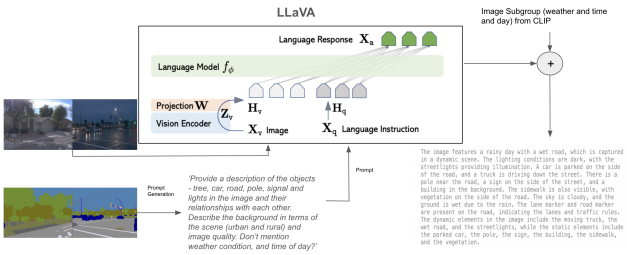

Figure 2. **Caption Generation Pipeline.** Our proposed caption generation scheme where we query a VLM such as LLAVA (image reproduced from the paper) with an image in the dataset, and a prompt constructed from the semantic masks to obtain a caption.

achieved through a mirrored U-Net encoder that guides image generation while maintaining the semantic mask's layout. When fine-tuning ControlNet for segmentation or autonomous driving, we freeze the LDM parameters. Training focuses on ControlNet's mirrored U-Net, which receives semantic masks. Text prompts are used to describe the image, incorporating the desired subgroup as a style. (See Sections 3.3 for details). We fine-tune ControlNet separately for each domain (segmentation, driving).

3.3. Improving text prompts via VLMs and Subgroups as style prompts for ControlNet

To improve synthetic data quality, we use high-quality captions as synthesis prompts. These captions, shown to enhance image realism [3, 13], are generated with LLaVA [22]. LLaVA leverages a VLM (Vicuna [7]) with CLIP-extracted image features to produce captions. In SynDiff-AD we generate captions for segmentation and driving datasets using LLaVA with a visual-questioning format. Queries are based on semantic mask classes, and we instruct LLaVA to exclude keywords related to target subgroups $z \in \mathcal{Z}$. This process is illustrated in Fig 2. Experiments in the Appendix demonstrate that LLaVA captions improve the realism of ControlNet synthesized images for segmentation tasks. Additionally, we use style prompting to ensure ControlNet captures the unique characteristics of each subgroup. Specifically, we augment the text prompts (generated as described in Section 3.3) with the subgroup z (identified by CLIP) and its semantic dimension. For example, if the dataset is partitioned by weather, we append "Image taken in Cloudy weather" to LLaVA's generated caption. This style prompting directs ControlNet to synthesize data tailored to specific under-represented subgroups.

3.4. Generating Synthetic Dataset

Synthesizing images with ControlNet for Segmentation and Autonomous Driving

We generate synthetic images for semantic segmentation using Algorithm 1. For autonomous driving, we make a key modification. Since driving annotations (e.g., waypoints) don't describe the visual scene, we need a semantic representation to ensure realistic synthesis. We obtain this layout either from a pre-trained segmentation model or, when available (like in the CARLA simulator), directly from provided semantic maps. Once we have the semantic layout, image synthesis proceeds as in the segmentation case.

Algorithm 1 Synthetic Data Augmentation Algorithm

Input: Dataset $D = \{(\mathbf{x}, \mathbf{y}, z)\}^N$ with images $\mathbf{x}$, masks $\mathbf{y}$, and annotated subgroups z, ControlNet $g_\theta(\mathbf{x}|\mathbf{y}, c)$ where c is a text caption, VLM Model - $VLM(\mathbf{x}, p)$ where p is a prompt, n_{synth} number of synthetic images
Output: Augmented Dataset D_{aug}
procedure DATAAUGMENTATION(D)
  $D_{\text{synth}} \leftarrow \{\}$
  for $i \leftarrow 1$ to n_{synth} **do**
    Sample $z \sim p_D(z)$ `// Sample subgroup from dataset`
    Sample $(\mathbf{x}, \mathbf{y}) \sim p_D(\mathbf{x}, \mathbf{y}|z)$ `// Sample image and mask given source subgroup from dataset`
    Obtain prompt p based on semantic label y and subgroup z
    $c \leftarrow VLM(\mathbf{x}, p)$ `// Caption c`
    Target subgroup $z_t \sim U(\mathcal{Z})$ `// Sample target subgroup uniformly`
    $c_t \leftarrow c \cup z_t$ `// Add target subgroup as style`
    Generate synthetic image $\mathbf{x}_{\text{synth}} \sim g_\theta(\mathbf{x}|\mathbf{y}, c_t)$
    $D_{\text{synth}} \leftarrow D_{\text{synth}} \cup \{(\mathbf{x}_{\text{synth}}, \mathbf{y})\}$
  end for
  $D_{\text{aug}} \leftarrow D \cup D_{\text{synth}}$
  return D_{aug}
end procedure

4. Results

Our experiments aim to analyze whether training models on synthetic data boosts overall segmentation and autonomous driving (AD) performance. The results for subgroup-specific performance improvements are mentioned in the Supplementary Material. We use two segmentation datasets: Waymo Open Dataset [34] (37,618 annotated images) and BDD100K [41] (7,000 annotated images). Our operational design domain includes weather (Rainy, Clear, Cloudy) and time of day (Dawn/Dusk, Morning, Night), yielding 9 subgroups (see Appendix). For AD, we use a CARLA simulator-based dataset with an expert driving policy [8]. Data is collected across 3 towns, 15 routes, with fixed weather per route for realism. Testing involves 27 routes under all weather conditions, using the same operational design domain as segmentation.

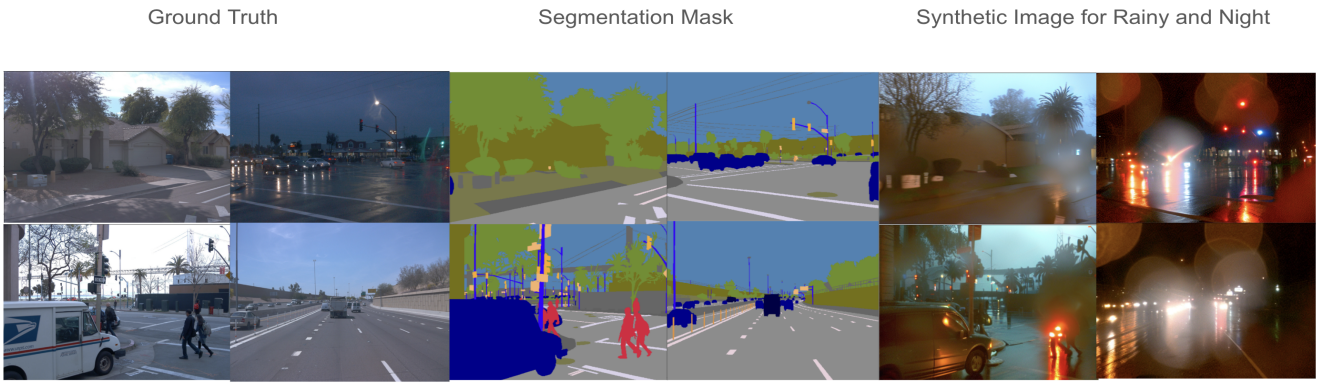

Figure 3. **Qualitative visualization of synthetic data for under-represented subgroups**.

Dataset	Model (Backbone)	Data	mIOU
Waymo	Mask2Former (R50)	Real	46.7
	Mask2Former (R50)	Real + Synth	47.8
	Mask2Former(Swin-T)	Real	50.2
	Mask2Former (Swin-T)	Real + Synth	**51.5**
	SegFormer (MIT-B3)	Real	45.4
	SegFormer (MIT-B3)	Real + Synth	46.2
BDD100K	Mask2Former (R50)	Real	51.5
	Mask2Former (R50)	Real + Synth	55.1
	Mask2Former (Swin-T)	Real	57.2
	Mask2Former (Swin-T)	Real + Synth	**58.6**
	SegFormer (MIT-B3)	Real	53.9
	SegFormer (MIT-B3)	Real + Synth	54.9

Table 1. **Models trained on augmented data outperform those trained on original data for segmentation tasks**.

Model	Data	Driving Score (↑)
NEAT [8]	Real	20.3
NEAT [8]	Real + Synth	13.63
AIM-2D [9]	Real	21.64
AIM-2D [9]	Real + Synth	34.36
AIM-BEV [9]	Real	28.61
AIM-BEV [9]	Real + Synth	**35.86**
Expert	-	54.6

Table 2. **Autonomous Driving performance improves with models trained augmented training datasets.**

4.1. Implementation Details

We finetune ControlNet on the Waymo, BDD (segmentation), and CARLA (autonomous driving) datasets. Both images and masks are resized to 512x512. We use a 5-epoch training regime, a learning rate of 1×10 5, and 8 A-5000 GPUs. During inference, Algorithm 1 guides style swapping as discussed earlier for synthetic image generation focused on under-represented subgroups. For semantic segmentation experiments we use Mask2Former (Swin-T and ResNet-50 backbones [14, 23]) and SegFormer (MIT-B3 backbone [38]. For E2E autonomous driving experiments, we finetuned NEAT, AIM-2D, and AIM-BEV [8] on a single A-5000 GPU for model training/evaluation.

4.2. Effect of Finetuning on final performance

We analyze segmentation models using standard metrics (mean IoU, mean Accuracy, mean DICE, mean F1), calculated across classes in the Waymo and BDD100K datasets. For autonomous driving (AD), we employ Route Completion, Infraction Score, and Driving Score [8, 9]. Synthetic data augmentation improves segmentation performance (Table 1). Notable gains include +1.3 mIoU for Mask2Former (Swin-T) on Waymo and +3.6 mIoU for Mask2Former (ResNet50) on BDD100K. In AD (CARLA simulator), AIM-2D and AIM-BEV benefit from synthetic data, primarily due to improved Infraction Scores (fewer collisions). NEAT, however, degrades, likely due to inconsistencies in spatial geometry within synthetic multi-camera views. We hypothesize that synthetic images provide context-sensitive augmentation, aiding generalization across visual and structural elements within subgroups.

5. Conclusion

This paper addresses the challenge of biased datasets that hinder autonomous driving performance in under-represented conditions. Our method SynDiff-AD provides a solution by leveraging advanced text-to-image generation with ControlNet to realistically augment datasets while preserving semantic details, eliminating the need for manual labeling. We showcase the capability of SynDiff-AD to improve segmentation and autonomous driving tasks. Future research could explore combining our generative approach with other data augmentation and task adversarial augmentation techniques and refining ControlNet to maintain spatial geometry within multi-camera views for AD. These advancements hold the potential to improve the robustness of AD systems significantly.

References

[1] Shubhankar Agarwal and Sandeep P. Chinchali. Synthesizing adversarial visual scenarios for model-based robotic control. In *Proceedings of The 6th Conference on Robot Learning*, pages 800–811. PMLR, 2023. 1

[2] Shubhankar Agarwal, David Fridovich-Keil, and Sandeep P Chinchali. Robust forecasting for robotic control: A game-theoretic approach. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5566–5573. IEEE, 2023. 1

[3] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023. 2, 3

[4] Dian Chen and Philipp Krähenbühl. Learning from all vehicles, 2022. 2

[5] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 2

[6] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 2

[7] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 3

[8] Kashyap Chitta, Aditya Prakash, and Andreas Geiger. Neat: Neural attention fields for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15793–15803, 2021. 2, 3, 4

[9] Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2, 4

[10] Krzysztof Czarnecki. Operational design domain for automated driving systems. *Taxonomy of Basic Terms ", Waterloo Intelligent Systems Engineering (WISE) Lab, University of Waterloo, Canada*, 2018. 2

[11] Fabio Henrique Kiyoiti dos Santos Tanaka and Claus Aranha. Data augmentation using gans, 2019. 2

[12] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[13] Shanghua Gao, Zhijie Lin, Xingyu Xie, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Editanything: Empowering unparalleled flexibility in image editing and generation. In *Proceedings of the 31st ACM International Conference on Multimedia, Demo track*, 2023. 2, 3

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[15] 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 preprint arXiv:2210.07574*, 2022. 2

[16] Yuru Jia, Lukas Hoyer, Shengyu Huang, Tianfu Wang, Luc Van Gool, Konrad Schindler, and Anton Obukhov. Dginstyle: Domain-generalizable semantic segmentation with image diffusion models and stylized semantic control, 2023. 2

[17] Abdulrahman Kerim, Felipe Chamone, Washington Ramos, Leandro Soriano Marcolino, Erickson R Nascimento, and Richard Jiang. Semantic segmentation under adverse conditions: a weather and nighttime-aware synthetic data-based approach. *arXiv preprint arXiv:2210.05626*, 2022. 1

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 2

[19] Shiye Lei, Hao Chen, Sen Zhang, Bo Zhao, and Dacheng Tao. Image captions are natural prompts for text-to-image models, 2023. 2

[20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 2

[21] Shaobo Lin, Kun Wang, Xingyu Zeng, and Rui Zhao. Explore the power of synthetic data on few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 638–647, 2023. 2

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 3

[23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2, 4

[24] Jan Hendrik Metzen, Robin Hutmacher, N Grace Hua, Valentyn Boreiko, and Dan Zhang. Identification of systematic errors of image classifiers on rare subgroups. *arXiv preprint arXiv:2303.05072*, 2023. 2

[25] Oskar Natan and Jun Miura. End-to-end autonomous driving with semantic depth cloud mapping and multi-agent. *IEEE Transactions on Intelligent Vehicles*, 8(1):557–571, 2022. 2

[26] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2022. 2

[27] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[28] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 1

[29] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics (TOG)*, 42(1):1–13, 2022. 2

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1, 2

[31] 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, 2023. 2

[32] Ibrahim Sobh, Loay Amin, Sherif Abdelkarim, Khaled Elmadawy, Mahmoud Saeed, Omar Abdeltawab, Mostafa Gamal, and Ahmad El Sallab. End-to-end multi-modal sensors fusion system for urban automated driving. 2018. 2

[33] Zhihang Song, Zimin He, Xingyu Li, Qiming Ma, Ruibo Ming, Zhiqi Mao, Huaxin Pei, Lihui Peng, Jianming Hu, Danya Yao, et al. Synthetic datasets for autonomous driving: A survey. *arXiv preprint arXiv:2304.12205*, 2023. 2

[34] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 1, 3

[35] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models, 2023. 2

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[37] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[38] 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*, 34:12077–12090, 2021. 2, 4

[39] Han Xue, Zhiwu Huang, Qianru Sun, Li Song, and Wenjun Zhang. Freestyle layout-to-image synthesis, 2023. 2

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

[41] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 1, 3

[42] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. 1, 2

[43] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort, 2021. 2

[44] Zheyuan Zhang, Lanhong Yao, Bin Wang, Debesh Jha, Elif Keles, Alpay Medetalibeyoglu, and Ulas Bagci. Emit-diff: Enhancing medical image segmentation via text-guided diffusion model. *arXiv preprint arXiv:2310.12868*, 2023. 2