WEDGE: A multi-weather autonomous driving dataset built from generative vision-language models

Aboli Marathe[1], Deva Ramanan[2], Rahee Walambe[3,4], Ketan Kotecha[3,4]

[1]Machine Learning Department, Carnegie Mellon University, PA
[2] Robotics Institute, Carnegie Mellon University, PA
[3] Symbiosis Centre for Applied AI (SCAAI), Symbiosis International University (SIU), India
[4]Symbiosis Institute of Technology (SIT), Symbiosis International University (SIU), India
abolim@cs.cmu.edu, deva@cs.cmu.edu, rahee.walambe@sitpune.edu.in, director@sitpune.edu.in

Abstract

The open road poses many challenges to autonomous perception, including poor visibility from extreme weather conditions. Models trained on good-weather datasets frequently fail at detection in these out-of-distribution settings. To aid adversarial robustness in perception, we introduce WEDGE (WEather images by DALL-E GEneration): a synthetic dataset generated with a vision-language generative model via prompting. WEDGE consists of 3360 images in 16 extreme weather conditions manually annotated with 16513 bounding boxes, supporting research in the tasks of weather classification and 2D object detection. We have analyzed WEDGE from research standpoints, verifying its effectiveness for extreme-weather autonomous perception. We establish baseline performance for classification and detection with 53.87% test accuracy and 45.41 mAP. Most importantly, WEDGE can be used to fine-tune state-of-the-art detectors, improving SOTA performance on **real-world** *weather benchmarks (such as DAWN) by* **4.48 AP for well-generated classes like trucks**. *WEDGE has been collected under OpenAI's terms [1] of use and is released for public use under the CC BY-NC-SA 4.0 license. The repository for this work and dataset is available at https://infernolia.github.io/WEDGE.*

1. Introduction

Self-driving cars need to safely operate across diverse weather conditions, generating a demand for extreme-weather perception data. This data is mostly captured through fleet operations which are dependent on several factors like sensor calibration, vehicle availability, road condition and equipment costs. Because of the low-frequency of naturally-encountered adverse weather, manual data collection can be expensive. Moreover, such collection can also

Figure 1. WEDGE synthetic images are generated from vision-language models using prompts of the form "{Objects} on {scenes} when {weather condition}". Crucially, weather conditions vary across {`snowing, raining, dusty, foggy, sunny, lightning, cloudy, hurricane, night, summer, spring, winter, fall, tornado, day, windy`}, as shown from the top-left to the bottom-right. By fine-tuning detectors on such images manually annotated with bounding boxes, we improve SOTA performance on real-world weather datasets [16] by **4.48 AP for well-generated classes like trucks.**

[1]https://openai.com/policies/terms-of-use

be unsafe for extreme weather conditions that reduce visibility or impair vehicle control, such as dust, snow, and fog. Because of such difficulty in data collection, many approaches treat weather conditions (such as rain droplets) as artifacts that can be removed through denoising [9, 21, 43].

One attractive alternative is the use of synthetic data built from rendering engines [29, 36], but such approaches may still not transfer to changing weather conditions or the realism of real-world, due to the so-called `sim2real` domain gap and underlying rendering assumptions. The recent development of realistic synthetic images with generative vision-language models (VLMs) suggests another approach: VLM prompting. We demonstrate that one can use VLMs to build adverse weather datasets for autonomous perception, improving performance on real-world datasets (such as DAWN [16]) for well-generated classes. Our main contributions include:

1. **Data.** First and foremost, we create WEDGE, a 3360 image synthetic dataset of autonomous driving scenes spanning 16 adverse weather conditions. We compare WEDGE to existing datasets, demonstrating that it includes more varied imagery.

2. **Release.** To allow for public release under the CC BY-NC-SA 4.0 license, we follow guidelines outlined by the VLM's terms of use, manual verifying the quality and appropriateness of the generated images.

3. **Annotation.** We provide ground truth annotations for all images for two tasks: weather classification and (2D) object detection, with, 16513 bounding box annotations.

4. **Benchmark.** We establish object detection and classification benchmarks, facilitating future work.

5. **Sim2real.** We provide initial evidence that suggests WEDGE can be used for sim2real learning; fine-tuning SOTA object detectors on WEDGE improves performance on real-world truck detection by **4.48** AP. We also examine object classes for which fine-tuning on WEDGE hurts performance.

The paper is organized as follows. Section 2 reviews prior datasets. Section 3 outlines the methodology used to construct and validate WEDGE. Section 4 presents experimental results for weather classification and sim2real object detection.

2. Background

The relationship between training data and test performance implies better generalization capabilities with better datasets. However, this assessment of "better" datasets can vary based on the respective task, expected performance, distribution requirements and other factors. In the context of autonomous driving tasks, we describe some recent datasets and the general requirements for robust models. We can see that as time progresses, larger datasets also expanded to include more weather conditions for robustness. However, even the best datasets till date do not venture beyond 4 weather types popularly.

Figure 2. **Real dataset samples from DAWN:** Weather conditions in the DAWN dataset [16] from top left to right: dust, fog, rainstorm, snowstorm.

Although a number of adverse weather datasets are reported in literature (Refer Table 1), they all pose limitations in two aspects : 1. The data contains the images corresponding to a very few (maximum four) adverse weather scenarios. 2. The data size is small, and it is biased towards a certain city or region and has an inter-class imbalance. When the models trained on these datasets are deployed for real-world computer vision tasks, their performance drops significantly due to lack of heterogeneity and variability. Hence, in this work, we report a new dataset which is developed using the DALL-E framework and offers balanced data generated for 16 weather scenarios and multiple object classes. The data is balanced for the all the weather events (210 images per weather class). The object class balance can also be achieved by weighting and re-sampling. Additionally, as the data is developed using generative AI, it is ideally more robust in nature. Some recent works have showcased favorable results using DALL-E and diffusion models for applications including zero-shot classification [18], detection [10] and face generation [3]. We provide a number of experimental results in support of robustness and evaluate the usability of this dataset as a benchmarking tool in autonomous perception.

3. Methodology

The dataset generation process, prompt formulation and image evaluation techniques are discussed here. The paper employs multiple analysis tools, frameworks, and models [2, 5, 8, 23, 26] to deliver the performance evaluation.

Work	Contribution	Features	Class Evaluated /Proposed	Cities	Weather Condition (S)
KITTI 2012 [11]	3D detection, stereo, optical flow, visual odometry/SLAM	22 scenes, stereo data, dense point clouds	3/3	1	Good weather only
CityScapes 2016 [7]	2D detection, semantic labeling	25000 images	19/30	50	Good weather only
Foggy Cityscapes Driving 2018 [32]	2D detection, semantic labeling	20,550 images	19/30	50	Fog
Waymo Open 2020 [37]	2D, 3D detection and tracking tasks	1150 scenes, LiDAR	4/4	3	Good weather with night, rain
nuScenes 2020 [4]	3D detection, tracking	1000 scenes, Radar data	10/23	2	Weather conditions (sun, rain and clouds)
DAWN 2020 [16]	2D detection	1000 scenes	6/6	-	Adverse weather: fog, snow, rain and sand
Argoverse 2 2023 [41]	3d tracking, motion forecasting	1000 scenes, HD maps	26/30	6	Weather include (sun, rain, snow)
WEDGE	2D Detection	3360 scenes	5/6	Unknown (variable)	Adverse weather in snowing, raining, dusty, foggy, sunny, lightning, cloudy, hurricane, night, summer, spring, winter, tornado, day, wind,fall

Table 1. Recent datasets in autonomous driving.

3.1. Ground-Truth Datasets

To test the weather durability of the zero-shot system, we set out to target a range of unfavorable weather situations that can degrade vision in any season. We need a benchmark poor-weather dataset from the actual world for a fair comparison in order to confirm the reliability of this dataset. The autonomous vehicle vision dataset: DAWN [16] with its 1000 driving scenarios recorded in adverse weather conditions was used for this test. Unfavorable weather conditions that are known to significantly limit road visibility include fog, snow, rain, tornadoes, haze, and sandstorms (Refer Fig 2). Bicycle, person (pedestrian), motorbike, truck, bus, and vehicle (car) form the set of 6 multiscale classes represented in the images.

3.2. WEDGE Dataset Generation

The DALL-E [28] is a large-scale text-to-image generation model that is based on an autoregressive transformer and has shown remarkable generalization capabilities in tasks like zero-shot learning. DALL-E 2 [27] is a dual-stage model that combines CLIP embeddings with probabilistic diffusion-model based decoder for conditional generation to generate the final realistic images. Diffusion models generate the images based on description (prompt) and sample using this condition. Due to the conditional generation, it presents the opportunity to generate variations in the generated images based on the embeddings. OpenAI has provided access to the DALL-E 2 model through OpenAI API which was used for dataset generation in the following steps:

1. Collected data using API calls to OpenAI API using prompts which were randomly sampled from the following sets of keywords:

   Scenes: highway, road, traffic jam, expressway Classes: cars, trucks, bus, people crossing Weather: snowing, raining, dusty, foggy, sunny, lightning, cloudy, hurricane, night, summer, spring, winter, fall, tornado, day, windy

2. Manually verified and cross-examined the images for errors, mismatch and inconsistencies.

3. Grouped images into categories based on weather keywords and thus generated 16 classes with 210 images of each class.

4. Generated 2D bounding box annotations for all images manually using RoboFlow annotation tool [8] and verified with human-in-the-loop evaluation.

5. Validated data using statistical and image analysis.

Specifically, we use prompts of the form "{Objects} on {scenes} when {weather}", where objects $\in$ {`cars, trucks, bus, people crossing`}, scenes $\in$ {`highway, road, traffic jam, expressway`}, and weather $\in$ {`snowing, raining,`

`dusty, foggy, sunny, lightning, cloudy, hurricane, night, summer, spring, winter, fall, tornado, day, windy`}. This is 4*4= 16 unique prompts for each weather condition, which we randomly queried 210 times to generate a final dataset of 16*210 = 3360 images. For the internal diagnostic analysis presented in Sec. 4, we randomly split WEDGE into a 80/20 train/test split for classification.

3.3. Image Similarity

We evaluate the threshold differences in image similarity between sampled real and generated images in their respective class clusters, and bin them as shown in Figure 5. The feature similarity index (FSIM) uses low level features to analyze images [44]. The Information theoretic-based Statistic Similarity Measure combines the statistical method and information theory, and it has a strong ability to forecast the relationship between the image intensity values [1]. Peak Signal-to-Noise Ratio (PSNR), which directly operates with image intensity, evaluates the ratio between the maximum possible power of a signal and the power of corrupting noise [13]. The Root Mean Squared Error (RMSE) calculates the percentage change in each pixel between the operation and the baseline [33]. Spectral Angle Mapper (SAM) calculates the angle between two spectra and treats them as vectors in a space with a dimensionality equal to the number of bands in order to estimate the spectral similarity between them [42].Signal to Reconstruction Error Ratio (SRE) is a metric that compares the error to the signal's power [17]. The Structural Similar Index Measure (SSIM) is a tool that aims to capture an image's loss of structure [13].

4. Experiments

4.1. Image Analysis

The classic autonomous vehicle settings contain skewed object distributions which we attempt to model with this generated dataset as visible in Figure 3. In practice, this balance can be restored by weighted prompting techniques and resampling if required, but should be maintained to deliver valid results benchmarking generalization capabilities.

We observe that the inter-class object distribution is also unbalanced (Figure 4), which is a desirable quantity for multi-weather robustness. In the wild, autonomous driving scenes will present unbalanced object distributions, which are difficult to perceive with detectors trained on fairly balanced data [25].

4.2. Image Similarity Analysis

We evaluate the real and generated datasets side by side using these 6 metrics and as seen in the figure 5, we hypothesize a sensible range of errors in this relative difference between real and generated datasets. The expected inverse similarity should ideally bounded by a small valued real number which varies with the similarity metric.

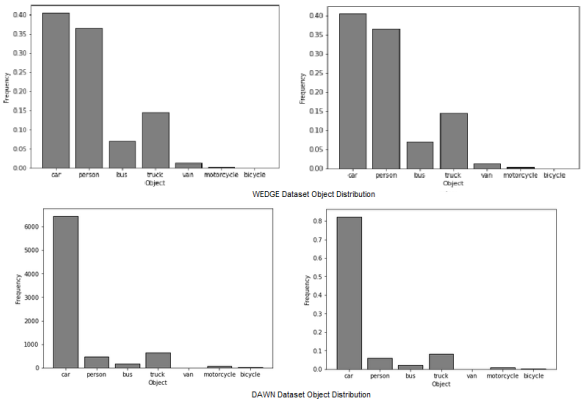

Figure 3. **Sim2Real Distribution Gap:** Object frequency distribution in WEDGE and DAWN datasets.

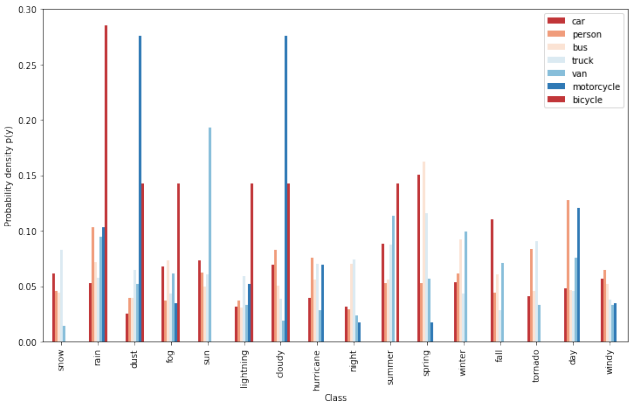

Figure 4. **Class Imbalance:** Inter-class object distribution in WEDGE dataset .

5. Results

5.1. Classification Benchmark

As visible in Table 2, the MobileNet [14] Classifier achieves top performance on the WEDGE Dataset with 53.87 test accuracy which is over 8-fold improvement on random classification that hits 6.25% accuracy.

5.2. Object Detection Benchmark

The main task of this study is examining for WEDGE's usefulness in robust object detection in multi-weather adversarial environments. Due to the varied presentations of results across previous works, we first establish a standard benchmark on DAWN dataset and attain dramatically better performance with 22.97 T-4 AP increase (on test set) and 17.07 T-4 AP increase (on complete set) (Table 3) than

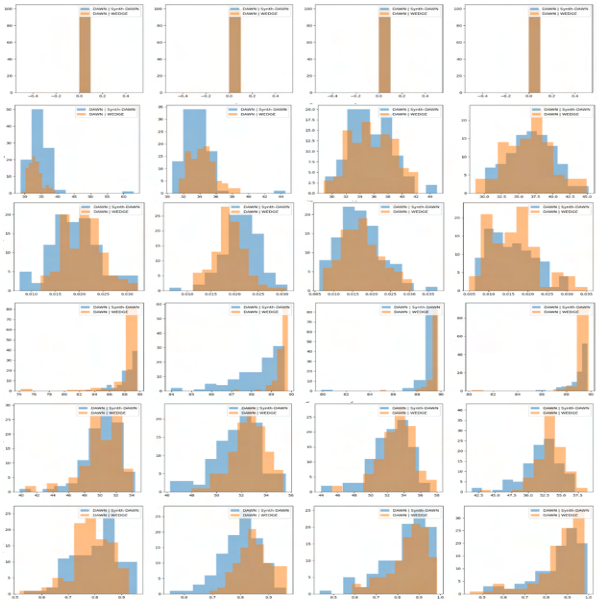

Figure 5. **Image similarity thresholds of modified real (Synth-DAWN) and synthetic autonomous driving datasets with real images as evaluated using 6 metrics (ISSM, PSNR, RMSE, SAM, SRE, SSIM from top to bottom) indicate overlapping similarity distributions**. The Sim2Real gap as evaluated by these metrics is comparable to Filter2Real gap of applying simple filters (like blurring, sharp edges without distorting the structural similarities). On the y-axis is the frequency of binned similarity(error) and x-axis is the bins of similarity. Orange color represents similarity between a sampled image from DAWN and WEDGE (Sim2Real). Blue color is similarity shift Filter2Real from common filtering modifications between a sampled image from DAWN and a modified sampled image from DAWN which is called Synth-DAWN.

Model	Train Acc.	Test Acc.
VGG16 [35]	95.80	44.35
VGG19 [35]	98.59	44.64
Xception [6]	98.33	46.73
ResNet50 [12]	35.04	22.47
ConvNeXtSmall [20]	58.97	18.15
InceptionV3 [38]	99.85	50.30
MobileNet [14]	99.33	**53.87**
MobileNetv2 [14]	99.67	46.43
DenseNet [15]	**99.89**	49.55
EfficientNetV2S [39]	35.01	18.75

Table 2. **Weather Classification**: Classifying the weather condition of a WEDGE image, when constructing a 80/20 train/test split. We see that models can predict weather conditions with reasonable accuracy.

the previous state-of-the-art ensemble works. While examining this performance, it is important to acknowledge that all previous works (Table 3) have evaluated on different DAWN test sets. For fair evaluation, we also publish our DAWN split (inclusive of all 4 weather conditions) and also evaluate on complete DAWN dataset. This is also a domain adaptation robustness benchmark contribution, as the models must shift between good-weather and adverse domains at test and training times respectively. We can see in Table 3, the FasterRCNN (ResNet 50) model fine-tuned on WEDGE achieves top detection performance for trucks, outperforming all previous benchmarks by 4.44 AP on truck object class for test set and 4.48 AP on trucks on complete set. The fine-tuned MobileNet (large) model is able to detect both cars and trucks better with 2.61 AP and 5.17 AP on test set (1.96 AP and 9.17 AP on complete set) respectively. We do not supervise the model with any images from DAWN in this training process. By fine-tuning on WEDGE (without access to DAWN data), we are able to attain multi-weather robustness in truck detection. However, one may question this robustness and attribute it to the large pre-training dataset used before fine-tuning. To verify the efficacy of WEDGE, we remove our fine-tuning module and demonstrate the performance of models. These models when trained without WEDGE (only good-weather data) are worse at detecting trucks. The most important insight from this work thus, is the effectiveness in utilizing WEDGE as a fine-tuning set and its importance for selective object detection. We observe that fine-tuning on WEDGE is insufficient for other classes due to inherent properties or annotation limitations. We manually examine how the synthetic objects in these class are significantly worse than real images which cause the detector to fine-tune on incorrect representations, thus hampering performance. Additionally we showcase the best object detection models in classical supervised settings attaining 45.41 mAP on the WEDGE Dataset with highest AP 57.48 on car class attained by Faster-RCNN (ResNet50).

6. Discussion

6.1. Qualitative Analysis

As shown in Fig. 1, we conduct qualitative analysis on generated samples and summarize our qualitative observations. `Snow` closely resemble winter scenes which contain noisy elements like snowfall and thus poor visibility. `Rain` resembles the view of a rainy traffic-filled road from the perspective of a sensor placed behind windshield. `Dust` contains occluded objects which are annotated for robust vision in adversity. `Fog` resembles dense foggy conditions which impair visibility of pedestrians and objects. `Sun` images have well-illuminated objects in variety of backgrounds. `Lightning` images look realis-

Model	Real Data (DAWN Dataset)								Synthetic Data (WEDGE Dataset)					
	car	person	bus	truck	T-4 AP	mc	bicycle	mAP	car	person	bus	truck	van	mAP
Benchmark														
Multi-weather city [24]	-	-	-	-	21.20	-	-	-	-	-	-	-	-	-
RoHL [31]	-	-	-	-	-	-	-	28.80	-	-	-	-	-	-
Transfer Learning [22]	7.00	8.00	7.00	-	5.50	-	0.00	-	-	-	-	-	-	-
Data Augmentation [22]	6.00	4.00	3.00	0.00	26.25	-	**92.00**	-	-	-	-	-	-	-
Weather-Night GAN [21]	48.00	0.00	0.00	0.00	12.00	-	-	-	-	-	-	-	-	-
Ensemble Detectors [40]	52.56	52.34	21.73	13.71	35.08	35.51	23.29	32.75	-	-	-	-	-	-
Evaluation on DAWN Test set														
Trained on Good Weather Data (COCO [19])														
FasterRCNN MobileNet Large 320 [14, 30]	39.08	22.71	37.13	10.78	27.42	8.33	0.00	19.70	34.10	36.26	39.35	16.05	0.00	25.15
FasterRCNN MobileNet Large [14, 30]	60.26	36.74	49.30	17.94	41.06	23.33	0.00	31.26	35.34	39.52	35.83	25.43	0.00	27.22
FasterRCNN ResNet 50 [30]	**71.19**	**69.51**	**69.88**	21.62	**58.05**	25.00	20.00	**46.20**	31.41	33.54	30.19	18.75	0.00	22.78
Fine-Tuning on WEDGE														
FasterRCNN MobileNet Large 320 [14, 30]	**41.69**	19.02	16.79	**15.95**	23.36	0.00	0.00	15.57	40.40	43.01	49.88	31.41	10.19	34.98
FasterRCNN MobileNet Large [14, 30]	58.54	28.39	29.14	**21.68**	34.43	0.00	0.00	22.96	52.52	**54.79**	**51.23**	50.01	7.95	43.30
FasterRCNN ResNet 50 [30]	65.47	39.70	54.19	**26.06**	46.35	0.00	0.00	30.9	**57.48**	54.71	46.92	**57.43**	**10.49**	**45.41**
Evaluation on Complete DAWN Dataset														
Trained on Good Weather Data (COCO [19])														
FasterRCNN MobileNet Large 320 [14, 30]	37.56	34.93	20.90	12.91	26.57	23.15	18.95	24.73	-	-	-	-	-	-
FasterRCNN MobileNet Large [14, 30]	60.64	55.96	32.78	23.66	43.26	38.55	28.75	40.05	-	-	-	-	-	-
FasterRCNN ResNet 50 [30]	**69.13**	**70.31**	**38.64**	30.54	**52.15**	**52.17**	**30.56**	**48.55**	-	-	-	-	-	-
Fine-Tuning on WEDGE														
FasterRCNN MobileNet Large 320 [14, 30]	**39.52**	23.97	7.81	**22.08**	23.34	0.00	0.00	15.56	-	-	-	-	-	-
FasterRCNN MobileNet Large [14, 30]	59.81	34.61	14.06	**30.67**	34.78	0.00	0.00	23.19	-	-	-	-	-	-
FasterRCNN ResNet 50 [30]	68.09	54.29	27.48	**35.02**	46.22	0.00	0.00	30.81	-	-	-	-	-	-

Table 3. **Object Detection:** Performance for `Car, Person, Bus, Truck, Van, Motorcycle (mc)` using the PASCAL VOC mAP metric on real (DAWN) and synthetic (WEDGE) datasets (90-10% , 0-100% train (unused)-test split of DAWN and 80- 16-4% train-val-test split for WEDGE for balanced comparison). First, we find that simply evaluating state-of-the-art (SOTA) object detectors (trained on *good* weather data) already outperforms all published results on DAWN. This establishes our pre-trained detectors as strong baselines for this task. Fine-tuning such models (specifically, ResNet50) on WEDGE further improves truck AP by 4.44 AP on test set (4.48 on complete set). The fine-tuned MobileNet-Large is able to detect both cars and trucks better with 2.61 AP and 5.17 AP on test set and (1.96 AP and 9.17 AP on complete set) respectively. T-4 AP is the averaged AP over the key object classes (car,person,bus,truck). Previous work 1 [24] was evaluated on DAWN WD set and reports AP. Previous work 2 [31] was evaluated on corrupted test sets in DAWN and reports mean AP over corruption types. Previous work 3, 4,5 [21, 22] was evaluated on 1000 images of DAWN and report AP and mAP. Previous work 6 [40] was evaluated on 500 images of DAWN and reports AP and mAP.

tic but typically contain a higher proportion of sky pixels. `Cloudy` resembles true cloudy scenarios with reduced illumination and gray overcasts. `Hurricane` consists of images that appear un-realistic, likely to due to the fact that this extreme weather condition is relatively rare. `Night` images have poor illumination and make detection difficult as expected. Often distant vehicles are just shown by blurred lights which we have included in annotations to ensure that vehicles can even detect distant mobile objects under low illumination. `Summer` are generally well-lit images.`Spring` images appear difficult to differenciate from day and sun, which is favorable as spring is a transitional season. `Winter` contains elements like snow, blizzards, hail which heavily obstruct vision and provide good adversaries to the detection task. Backgrounds are mostly white and snow–covered which makes the detection task simpler. This does not represent winter in warmer countries, which must be treated by mixing classes. `Fall` im-

ages are skewed to geographic regions that are usually associated with the aesthetic fall backgrounds including bright trees, fallen leaves which are mostly present in the northern regions of countries. `Tornado` contains a good number of unrealistic images as well, but manages to capture the essence of this natural disaster through poor-illumination, windy conditions and distant tornado funnels. In the unrealistic cases, tornadoes appear in extremely unlikely scenarios like exactly on top of the car, as visualized in cartoons and games. `Day` images are well-lit and show sunny scenarios, also including some overcast skies. `Windy` images are either realistic or extremely skewed towards disaster-like scenarios including uprooting winds, destroyed vehicles and fading objects.

As visible in Fig. 6 we highlight some possible causes of poor performance. Region-centric correlations (eg. cherry blossoms associated with spring), are a recurring theme in the generated images inspite of providing generic prompts. Generative anomalies like extra terrestrial creatures crossing the road occur when the terrain described in prompt (dust) matches similar out-of-distribution examples (Martian imagery). Training objects sometimes combine to form interesting but unrealistic characters in this synthetic in spite of given realistic prompts. We also identify entities with missing parts (humans without heads). While this feature can help improve robustness to occlusion, it is still a defect in generated images. Often typical scenes which correspond to the prompt are generated in sketch, animated or miscellaneous styles. Objects which are closer to the viewer (camera)'s supposed location are more accurately generated. As seen in the figure and other examples, the distant objects are often lacking quality and fundamental differentiating characteristics which are necessary for detectors. Although we cannot accurately pinpoint the time frame of generated images, we observe special cases of people wearing masks in locations (predicted locations) where masks were not worn prior to the pandemic. While this may be attributed to different reasons, we can consider this feature as an important part of robustness in post-pandemic systems. As the prompts shift to more out-of-distribution settings, like tornadoes, we observe a dramatic shift in favor of unrealistic images. This may be due to the inability to find hyper-realistic training images captured in these adverse conditions, but are a potential limitation. Spatial anomalies frequently distort the placement, positioning, orientation and interaction between generated objects. In this case, we observe shadows are generated inconsistently. As generative models move closer towards real-world simulation, focus on modeling relationships between entities on the basis of physical, scientific and behavioral properties can be explored. Beneficial anomalies like scenes generated around accidents, mishaps like tire punctures, car crashes and weather-related disasters like tornadoes uprooting the roofs of buses appear often in the data. These accidents are very realistic and not often captured by common autonomous vehicle datasets. These scene-specific datasets can be generated for detecting emergencies in surveillance systems. Human generation ultimately presents the greatest challenge in dataset utilization. The images of humans in the dataset have second-largest frequency but are often unrealistic (either to the out-of-distribution prompts or intentional obscuring done for privacy concerns) which can potentially affect fine-tuning as seen in the previous section.

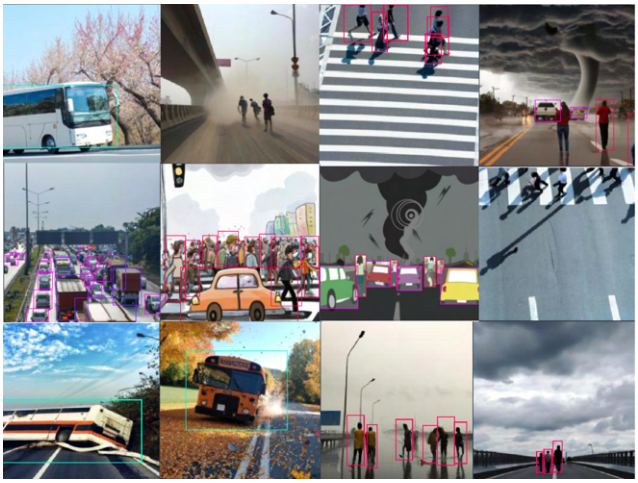

Figure 6. **Qualitative Analysis:** Limitations of WEDGE appear in the form of region-centric spurious correlations, generative anomalies,missing (incomplete generation) features,domain and style transfer, distance (proximity to viewing angle) bias, multi-time relevance, class (weather) bias, spatial and placement anomalies,human generation inconsistencies from top-left to bottom-right.

6.2. Benefits of generated datasets (WEDGE)

The feature and capability of embedding variability from vast text corpora into image sets using prompting of generative models provides support for building robust models. Due to the high variability in geography, population, seasons, weather, illumination, perspective and backgrounds, models are able to generalize to detect trucks on real roads. We can simulate specific out-of-distribution scenarios like road accidents to monitor safety through anomaly detection or other tasks.

6.3. Potential limitations of generated datasets (WEDGE):

Although the variability is assumed to be limited to the prompt space, it is practically not constrained by the prompt constraints. We observe the generation going beyond realistic scenes and stepping into style transfers, anomaly generations and overall extreme variability which is not always

required. This generation will be based on the training data of the large-scale generative model, which has been treated with algorithmic changes by model creators, but can be improved. Additional bias is propagated due to prompting styles, language and keywords used. In general, we observe that the generated images report some consistency problems with respect to physical properties and orientation, spatial compatibility and interaction between the objects. Poor generation can result in certain properties like leaking lights, erroneous green patches and paint-like patches on random images. The most important result is the frequency of poorly generated objects directly impacts the resultant performance on real-images which should be considered in future work.

6.4. When does WEDGE work?

The image generation procedure and results of this study speak in favor of its importance to the autonomous driving perception tasks. Prompting was focused on generating the most relevant autonomous vehicle-related images for 16 weather-classes and manually verified. Image screening and curation was performed to ensure inter-class-prompt consistency. The provided annotations and extensive bounding boxes (16513) for all classes have been generated with human-in-the-loop. 16 unique weather-seasonal variations captured for autonomous vehicles which is unique to this dataset and essential for multi-weather robustness. Annotations for heavily occluded and obscure objects (headlights in fog) have been labelled to assist models in learning representations from occluded objects. Inspite of having out-of-distribution scenarios, the image similarity thresholds are still within reasonable range from the sample distribution shifts which speaks in favor of data adoption in similar tasks needing sensor-based data. Models trained on WEDGE for domain adaptive detection were able to cross the benchmark on the DAWN dataset in under-represented target classes like trucks. The difference between generated people and trucks are their similarities to real-world objects which differ dramatically between real and synthetic data, thus offering a plausible explanation for the performance difference.

7. Conclusion

In this work, we explore AI-generated datasets [2] for robust multi-weather perception. We perform a small-scale analysis of its task-specific properties in the context of autonomous vision and demonstrate the effectiveness of such generation. Under the constraints of selected data, we assess the usefulness of these datasets from the perspective of autonomous perception. We acknowledge that all findings are constrained to this case study between the selected domain and target data only, and do not present findings for perception or synthetic data in general. In this work, we additionally present a state-of-the-art benchmark for DAWN dataset using standard evaluation metrics (without any access to target or adverse-weather training data) through robust performance evaluation. We hope to aid in the effort towards meeting the need for autonomous vision datasets by this demonstration. In future works, this data generation procedure paired with creative prompt engineering can deliver superior performance in multi-weather domains.

[2] All references to "generated" in this text imply AI generated datasets only. The authors generated this dataset in part with DALLE-2, OpenAI's large-scale image-generation model. Upon generating the dataset, the authors reviewed the images and take responsibility for their content in accordance with the terms laid out by OpenAI. The authors have created "Input" prompts on their own and obtained data "Output" images **only** using the official OpenAI API through a paid subscription service.

Figure 7. **WEDGE as an adversarial example:** We observe significant shifts in attention maps [34] when data contains poor-weather conditions. The object of interest was the vehicle in the images which the attention maps are not following due to the weather-based corruptions of fog and dust. This provides support to why good-weather data are often insufficient while building robust perception models.

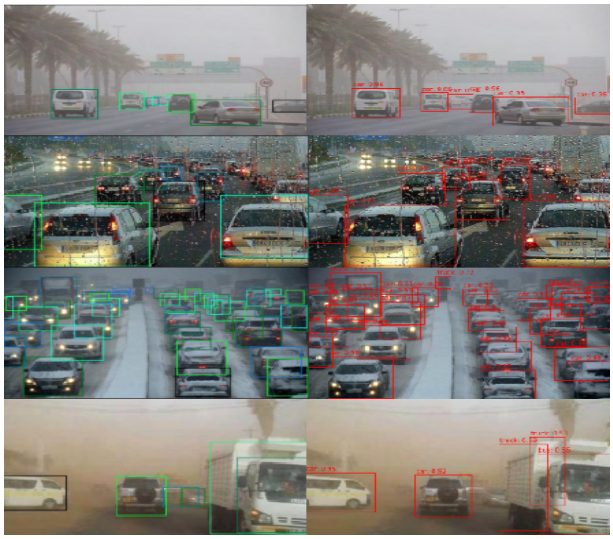

Figure 8. **Sim2Real Inference:** Comparison of (COCO) pre-trained Resnet 50 Faster RCNN (**left**) with a variant fine-tuned on WEDGE (**right**) on a test image from DAWN. We see that the fine-tuned models tend to predict trucks better but suffer from false positives, resulting in lower car APs.

References

[1] Mohammed Abdulameer Aljanabi, Zahir M Hussain, Noor Abd Alrazak Shnain, and Song Feng Lu. Design of a hybrid measure for image similarity: a statistical, algebraic, and information-theoretic approach. *European Journal of Remote Sensing*, 52(sup4):2–15, 2019. 4

[2] Alan Bi. Welcome to detecto's documentation!, Accessed on 6 April 2023. 2

[3] Ali Borji. Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2. *arXiv preprint arXiv:2210.00586*, 2022. 2

[4] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 3

[5] Joao Cartucho, Rodrigo Ventura, and Manuela Veloso. Robust object recognition through symbiotic deep learning in mobile robots. In *2018 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 2336–2341. IEEE, 2018. 2

[6] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 5

[7] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 3

[8] B Dwyer and J Nelson. Roboflow (version 1.0). *URL https://roboflow. com*, 2022. 2, 3

[9] David Eigen, Dilip Krishnan, and Rob Fergus. Restoring an image taken through a window covered with dirt or rain. In *Proceedings of the IEEE international conference on computer vision*, pages 633–640, 2013. 2

[10] Yunhao Ge, Jiashu Xu, Brian Nlong Zhao, Laurent Itti, and Vibhav Vineet. Dall-e for detection: Language-driven context image synthesis for object detection. *arXiv preprint arXiv:2206.09592*, 2022. 2

[11] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 3

[12] 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. 5

[13] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 4

[14] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 4, 5, 6

[15] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 5

[16] Mourad A Kenk and Mahmoud Hassaballah. Dawn: vehicle detection in adverse weather nature dataset. *arXiv preprint arXiv:2008.05402*, 2020. 1, 2, 3

[17] Charis Lanaras, José Bioucas-Dias, Silvano Galliani, Emmanuel Baltsavias, and Konrad Schindler. Super-resolution of sentinel-2 images: Learning a globally applicable deep neural network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 146:305–319, 2018. 4

[18] Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. *arXiv preprint arXiv:2303.16203*, 2023. 2

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[20] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 5

[21] Aboli Marathe, Pushkar Jain, Rahee Walambe, and Ketan Kotecha. Restorex-ai: A contrastive approach towards guiding image restoration via explainable ai systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3030–3039, 2022. 2, 6

[22] Aboli Marathe, Rahee Walambe, Ketan Kotecha, and Deepak Kumar Jain. In rain or shine: Understanding and overcoming dataset bias for improving robustness against weather corruptions for autonomous vehicles. *arXiv preprint arXiv:2204.01062*, 2022. 6

[23] Markus U Müller, Nikoo Ekhtiari, Rodrigo M Almeida, and Christoph Rieke. Super-resolution of multispectral satellite images using convolutional neural networks. *arXiv preprint arXiv:2002.00580*, 2020. 2

[24] Valentina Muşat, Ivan Fursa, Paul Newman, Fabio Cuzzolin, and Andrew Bradley. Multi-weather city: Adverse weather stacking for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2906–2915, 2021. 6

[25] Kemal Oksuz, Baris Can Cam, Sinan Kalkan, and Emre Akbas. Imbalance problems in object detection: A review. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3388–3415, 2020. 4

[26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit

Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 2

[27] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[28] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 3

[29] William T Reeves. Particle systems—a technique for modeling a class of fuzzy objects. *ACM Transactions On Graphics (TOG)*, 2(2):91–108, 1983. 2

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 6

[31] Tonmoy Saikia, Cordelia Schmid, and Thomas Brox. Improving robustness against common corruptions with frequency biased models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10211–10220, 2021. 6

[32] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018. 3

[33] Umme Sara, Morium Akter, and Mohammad Shorif Uddin. Image quality assessment through fsim, ssim, mse and psnr—a comparative study. *Journal of Computer and Communications*, 7(3):8–18, 2019. 4

[34] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 8

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[36] Karl Sims. Particle animation and rendering using data parallel computation. In *Proceedings of the 17th annual conference on Computer graphics and interactive techniques*, pages 405–413, 1990. 2

[37] 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. 3

[38] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 5

[39] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021. 5

[40] Rahee Walambe, Aboli Marathe, Ketan Kotecha, George Ghinea, et al. Lightweight object detection ensemble framework for autonomous vehicles in challenging weather conditions. *Computational Intelligence and Neuroscience*, 2021, 2021. 6

[41] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023. 3

[42] Roberta H Yuhas, Alexander FH Goetz, and Joe W Boardman. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. In *JPL, Summaries of the Third Annual JPL Airborne Geoscience Workshop. Volume 1: AVIRIS Workshop*, 1992. 4

[43] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022. 2

[44] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011. 4