Multiple Distribution Shift - Aerial (MDS-A): A Dataset for Test-Time Error Detection and Model Adaptation

Noel Ngu[1], Aditya Taparia[1], Gerardo I. Simari[2], Mario Leiva[2], Ransalu Senanayake[1], Paulo Shakarian[1], Nathaniel D. Bastian[4], John Corcoran[3]

[1]Arizona State University, Tempe, AZ USA
[2]Department of Computer Science and Engineering, Universidad Nacional del Sur and Institute for Computer Science and Engineering, Bahía Blanca, Argentina
[3]U.S. Department of Defense, Arlington, VA USA
[4]United States Military Academy, West Point, NY USA
nngu2@asu.edu, ataparia@asu.edu, gis@cs.uns.edu.ar, mario.leiva@cs.uns.edu.ar, jack.fd.corcoran@gmail.com, ransalu@asu.edu, pshak02@asu.edu, nathaniel.bastian@westpoint.edu

Abstract

Machine learning models assume that training and test samples are drawn from the same distribution. As such, significant differences between training and test distributions often lead to degradations in performance. We introduce Multiple Distribution Shift - Aerial (MDS-A) - a collection of inter-related datasets of the same aerial domain that are perturbed in different ways to better characterize the effects of out-of-distribution performance. Specifically, MDS-A is a set of simulated aerial datasets collected under different weather conditions. We include six datasets under different simulated weather conditions along with six baseline object-detection models, as well as several test datasets that are a mix of weather conditions that we show have significant differences from the training data. In this paper, we present characterizations of MDS-A, provide performance results for the baseline machine learning models (on both their specific training datasets and the test data), as well as results of the baselines after employing recent knowledge-engineering error-detection techniques (EDR) thought to improve out-of-distribution performance. The dataset is made readily available online.

Datasets — https://lab-v2.github.io/mdsa-dataset-website

Introduction

The robustness of models for object-detection remain a critical challenge when dealing with distributional shifts in real-world data. Distributional shifts in weather are especially important in aerial imagery since visibility and object-recognition can be heavily influenced by the weather. Prior work on establishing benchmarks for out-of-distribution (OOD) object detection has largely focused on evaluating existing model performance during such a shift (Mao et al. 2023; Gardner, Popović, and Schmidt 2024). In this work, we present the Multiple Distribution Shift - Aerial (MDS-A) dataset - a collection of generated and labeled datasets with varying distribution differences and an associated set of baseline models. To control experiments, we keep the baseline domain (aerial imagery) constant and perturb it in different ways to better characterize the effects of out-of-distribution performance. Specifically, MDS-A is a set of simulated aerial datasets taken under different weather conditions. We include six datasets under different simulated weather conditions along with six baseline object detection models as well as several test datasets that are a mix of weather conditions that we show have significant differences from the training data. In this paper, we present characterizations of MDS-A, provide performance results for the baseline models (on both in-distribution and out-of-distribution test sets), as well as results of the baselines after employing recent knowledge-engineering error-detection techniques (error detection rules, or EDR (Kricheli et al. 2024; Xi et al. 2024; Lee et al. 2024; Shakarian, Simari, and Bastian 2025)) thought to improve out-of-distribution performance. The rest of the paper is organized as follows. First, we introduce the dataset, describing how it was constructed, and reporting on key statistics, importantly measures of distributional differences between the various training and testing sets. Then, we describe how we trained a series of baseline models, and report on their performance both with and without error detection rules. Finally, we discuss future research directions in the conclusion.

Dataset

In this section, we describe how we created the MDS-A dataset and report key statistics including measures of distributional differences.

AirSim simulator To investigate the impact of distributional shift on aerial imagery in the context of weather conditions, we employed AirSim, an open-source simulator for drones, ground vehicles, cars, and other objects (Shah et al. 2018), to create a dataset of aerial imagery under various weather conditions. AirSim provides tools to capture images from different positions under different weather conditions by adjusting configurable parameters for effects such as dust, rain, fog, snow, and maple leaves. These parameters give us control over the intensity of various weather effects in the simulated scenes. Panel A and B in Figure 1 demonstrates how changing these parameters visually impact the images captured in AirSim.

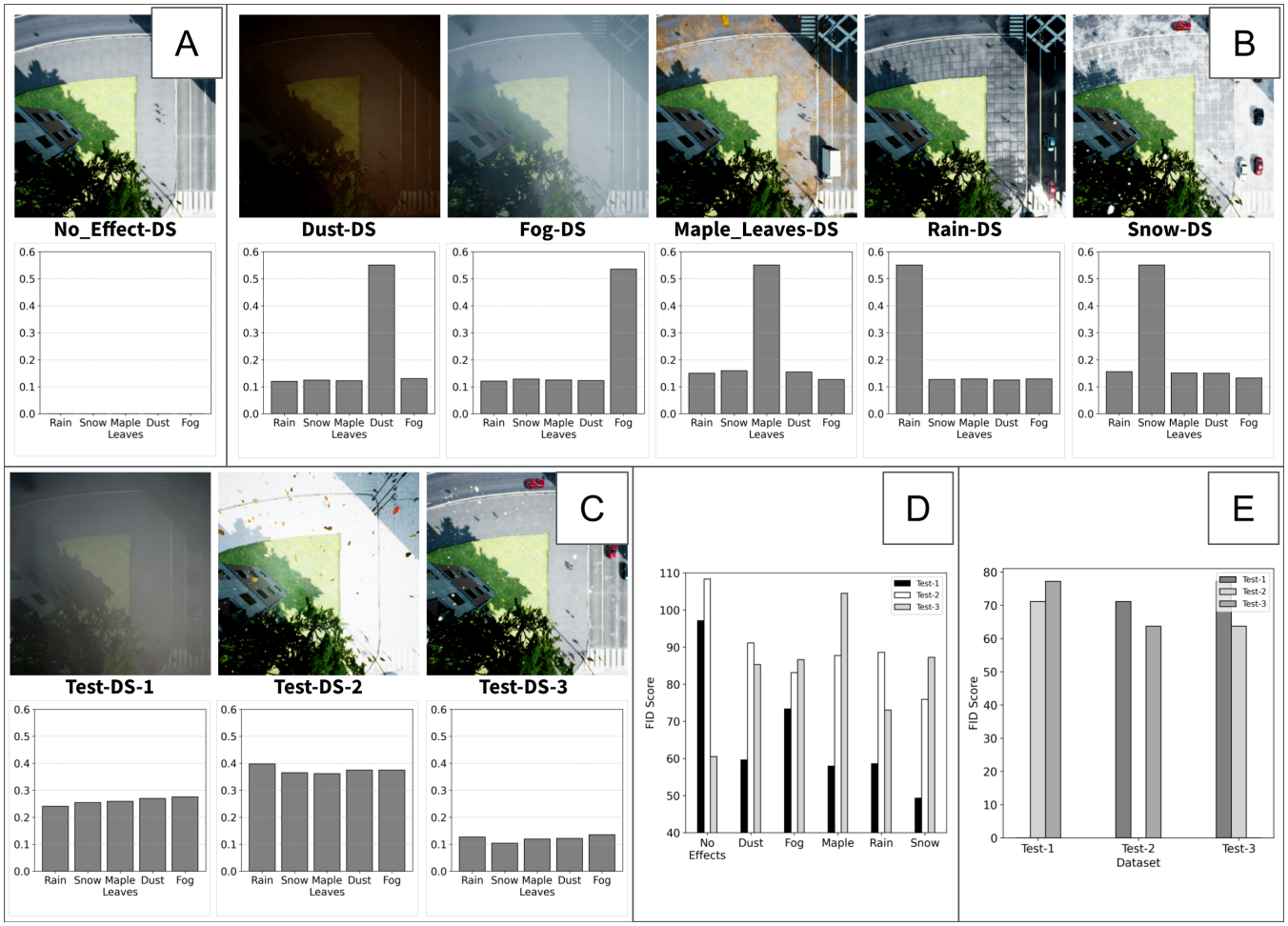

Figure 1: A) Image captured in AirSim with no weather effects applied along with a histogram showing the distribution of weather conditions of the dataset that it represents. B) Images captured in the same position in AirSim under different weather conditions: dust, fog, maple leaves, rain, snow-along with a histogram showing the distribution of weather conditions of the dataset that it represents. C) Images captured in the same position in AirSim with a mix of weather conditions applied along with a histogram showing the distribution of weather conditions of the dataset that it represents. D) A histogram showing the FID scores between the training sets and the 3 test sets. E) A histogram showing the FID score comparisions between the test sets.

Data collection For this study, the drone vehicle in AirSim was utilized to capture images from a top-down view at random positions within a simulated city environment. Given evidence that state-of-the-art object detection models are often susceptible to diverse weather conditions (Pathiraja, Liu, and Senanayake 2024), we configured AirSim with the following weather effects: rain, snow, fog, maple leaves, and dust. Using AirSim, images along with their bounding boxes were generated. Objects in the captured scenes were labeled by the research team to be classified into the following four categories: pedestrians, vehicles, nature, and construction. Each bounding box was assigned to exactly one of the categories.

- **Training sets** Multiple training sets were created, each focusing on a specific weather condition. For each dataset, the corresponding weather parameter for a weather condition (e.g.rain, snow, fog, maple leaves, or dust) are set randomly with a specific weather condition set to a particularly high value, while the other weather parameters were set to low values. As a result, we created distinct training datasets for the following conditions: Rain, Snow, Fog, Maple leaves, Dust. Panel B in Figure 1 shows the average intensities of each weather condition in each training set. In addition, a training set with no weather effects was created as well. The objective of these training sets are to enable the models to specialize in identifying objects under a single dominant weather condition.
- **Test sets** The test set was designed to evaluate the ability of models (trained on the training datasets) on a dataset that was created to simulate natural distributional shifts in weather. Unlike the training set, the test set consists

Name	Images	Bounding boxes
No-Effect Train Set	1000	13320
Dust Train Set	1000	11257
Fog Train Set	1000	11099
Maple-Leaves Train Set	1000	12295
Rain Train Set	1000	11528
Snow Train Set	1000	11462
No-Effect Test Set	100	1267
Dust Test Set	100	1255
Fog Test Set	100	1406
Maple-Leaves Test Set	100	1077
Rain Test Set	100	1299
Snow Test Set	100	1368
Test Set 1	1000	11117
Test Set 2	1000	12466
Test Set 3	1000	12558

Table 1: Statistics regarding the number of images and the number of bounding boxes in each training set and test set.

of complex weather conditions where multiple weather conditions could be set to high values simultaneously, creating more challenging object-detection samples for the models. Panel C in Figure 1 shows the average intensities of each weather condition in the test set.

Dataset Statistics MDS-A consists of training sets that focus on a single weather condition, with each set containing 1000 images. The following training sets were generated: No-Effect Train Set, Dust Train Set, Fog Train Set, Maple-Leaves Train Set, Rain Train Set, Snow Train Set. The test sets, in contrast, feature a complex combination of weather conditions, also comprising 1000 images. Table 1 shows some statistics regarding the number of images and the number of bounding boxes in each training set and test set. We note that with each of the six training sets, there is also a corresponding in-distribution hold-out set containing 100 images- this allows us to compare model in-distribution performance with out-of-distribution performance easily, all the while controlling for other factors.

Additionally, the Fréchet Inception Distance (FID) (Heusel et al. 2018) scores between the training sets and the test set are presented in Panel D in Figure 1. These scores reflect the visual similarity between the training sets and the test set, providing a way to approximate the amount of distributional-shift between the training set and the test set. Higher FID scores, especially for conditions like Fog (73.3), suggests a larger distributional shift between the training set and the test set.

Metadata Conditions In addition to the datasets, we also provide additional meta conditions for each sample. This information can be used to learn metacognitive models to identify potential errors. We use these in our baselines for error detection later in the paper. Examples of such conditions can be seen in Table 2

Rule	Meaning of Rule
$cond_{green}(w)$	Colors inside the bounding box has to be green
$cond_{overlap}(w)$	Pedestrians and vehicles should not overlap.

Table 2: Example EDCR Rule Learned for the MPSC Problem

Baseline Models and Associated Performance

In addition to providing a dataset, we provide a series of baseline models, in addition to employing error detection rules (Kricheli et al. 2024; Xi et al. 2024; Lee et al. 2024; Shakarian, Simari, and Bastian 2025).

Model Training In order to establish a baseline for model performance under distributional shifts in the context of weather conditions, object-detection models were trained on each training set. The baseline object detection model that was used was DeTR (Carion et al. 2020) with a ResNet-50 (He et al. 2016) backbone.

The models were intentionally trained on a single training set without any mixes between training sets in order to emphasize different weather effects. These models were then evaluated on a more complex dataset aimed to emulate natural distributional shifts in weather conditions.

In-Distribution Model Performance Table 4 provides results of the baseline models on their corresponding in-distribution dataset, specifically the *No Effect Test Set*, *Dust Test Set*, *Fog Test Set*, *Maple-Leaves Test Set*, *Rain Test Set*, and *Snow Test Set* (see statistics in Table 1 for details). Here we report precision, recall, and F1 (harmonic mean of precision and recall). We note that model performance is generally consistent across the various models.

Performance of Baseline Models on Test Sets The baseline models were evaluated on out-of-distribution test sets to assess their robustness under complex weather conditions that differ from the distribution in which they were trained - this is to establish a baseline for out-of-distribution performance on the three test sets in MDS-A. Table 3 shows an expected decline in precision, recall, and F1 compared to in-distribution results.

Models with Error Detection Rules To enhance the robustness of the baseline models, error detection rule learning (EDR) was applied using the DetRuleLearn algorithm (Xi et al. 2024) with the hyperparameter of ϵ set to 0.5. Note that the rules were trained on the same data as the models.v The application of EDR showed improvements in Precision while mostly maintaining F1 across all test sets as shown in Table 3. This is due to the fact that EDR rules produce detections that are essentially recognizing that the model will most likely produce an error - and hence the results are discarded - resulting in a reduction of recall but an increase in precision. We note that the results of (Xi et al. 2024) associate recall reduction with the ϵ hyperparameter (which would be up to an 0.5 reduction, see Theorem 2 in (Xi et al. 2024)) - however it is noteworthy that the reduction in recall is much less than predicted by the theoretical guarantee.

Model	Precision	Recall	F1	Precision (EDR)	Recall (EDR)	F1 (EDR)
Test Set 1						
No Effect Model	0.35	0.27	0.31	**0.62**	0.25	0.36
Snow Model	0.59	**0.55**	**0.57**	0.61	0.50	0.55
Dust Model	0.59	0.54	**0.57**	0.61	0.49	0.54
Maple Leaf Model	0.60	**0.55**	**0.57**	0.60	**0.55**	**0.57**
Rain Model	0.60	0.54	**0.57**	0.60	0.54	**0.57**
Fog Model	0.56	0.53	0.55	0.56	0.53	0.55
Test Set 2						
No Effect Model	0.16	0.14	0.15	**0.54**	0.13	0.21
Snow Model	0.44	**0.26**	**0.32**	0.47	0.25	**0.32**
Dust Model	0.43	0.25	**0.32**	0.46	0.24	**0.32**
Maple Leaf Model	0.45	0.25	**0.32**	0.45	0.25	**0.32**
Rain	0.46	0.25	**0.32**	0.46	0.25	**0.32**
fog	0.40	0.25	0.31	0.40	0.25	0.31
Test Set 3						
No Effect Model	0.50	0.35	0.41	**0.65**	0.30	0.41
Snow Model	0.63	0.52	**0.57**	**0.65**	0.49	0.56
Dust Model	0.58	0.47	0.52	0.60	0.43	0.50
Maple Leaf Model	0.61	**0.53**	**0.57**	0.61	**0.53**	**0.57**
Rain Model	0.57	0.47	0.52	0.57	0.47	0.52
Fog Model	0.55	0.42	0.48	0.55	0.42	0.48

Table 3: Table showing the before and after results of applying EDR. Underlined numbers indicates the best model. Bold numbers indicates the best performing model across both baseline and EDR.

Model	Precision	Recall	F1
No Effect	0.75	0.62	0.68
Snow	0.75	0.69	0.72
Dust	0.75	0.65	0.70
Maple Leaves	0.76	0.70	0.73
Rain	0.75	0.65	0.70
Fog	0.73	0.62	0.67

Table 4: Table showing the performance of the baseline models trained on different training sets on an in-distribution dataset that is distinct from the training set.

Conclusion and Future Work

In this paper, we introduced the Multiple Distribution Shift - Aerial (MDS-A) dataset, a collection of simulated aerial datasets made to investigate the impact of distributional shifts, in the context of weather conditions, on object-detection model performance. Using the AirSim simulator, we created training datasets under six distinct weather conditions—rain, snow, fog, maple leaves, dust, and no effects—and evaluated the performance of baseline object-detection models trained on each condition using a complex test set that combines multiple weather effects. We also provide a suite of baseline models and in this paper we report on their performance for both in-distribution and out-of-distribution datasets. Additionally, we also provide a baseline using error detection rules, which mitigates the degradation of precision. As we intend this to be a challenge dataset, we released MDS-A and the associated baseline models at https://lab-v2.github.io/mdsa-dataset-website.

Recent advances in topics such as test time training (Liang, He, and Tan 2025), domain generalization (Zhou et al. 2023), and meta learning (Vanschoren 2018) are all potential candidates for improving performance. Further, this dataset allows the exploration of novel ensemble methods based on models trained on different distributions.

Acknowledgments

This research was supported by the Defense Advanced Research Projects Agency (DARPA) under Cooperative Agreement No. HR00112420370, the U.S. Army Combat Capabilities Development Command (DEVCOM) Army Research Office under Grant No. W911NF-24-1-0007, and the U.S. Army DEVCOM Army Research Lab under Support Agreement No. USMA 21050. The views expressed in this paper are those of the authors and do not reflect the official policy or position of the U.S. Military Academy, the U.S. Army, the U.S. Department of Defense, or the U.S. Government.

Thank you for reading these instructions carefully. We look forward to receiving your electronic files!

References

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, 213–229. Berlin, Heidelberg: Springer-Verlag. ISBN 9783030584511.

Gardner, J.; Popović, Z.; and Schmidt, L. 2024. Benchmarking distribution shift in tabular data with TableShift. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, 53385–53432. Red Hook, NY, USA: Curran Associates Inc.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2018. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. (arXiv:1706.08500). ArXiv:1706.08500.

Kricheli, J. S.; Vo, K.; Datta, A.; Ozgur, S.; and Shakarian, P. 2024. Error Detection and Constraint Recovery in Hierarchical Multi-Label Classification without Prior Knowledge. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, 3842–3846. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704369.

Lee, N.; Ngu, N.; Sahdev, H. S.; Motaganahall, P.; Chowdhury, A. M. S.; Xi, B.; and Shakarian, P. 2024. Metal Price Spike Prediction via a Neurosymbolic Ensemble Approach. (arXiv:2410.12785). ArXiv:2410.12785.

Liang, J.; He, R.; and Tan, T. 2025. A Comprehensive Survey on Test-Time Adaptation Under Distribution Shifts. *International Journal of Computer Vision*, 133(1): 31–64.

Mao, X.; Chen, Y.; Zhu, Y.; Chen, D.; Su, H.; Zhang, R.; and Xue, H. 2023. COCO-O: A Benchmark for Object Detectors under Natural Distribution Shifts. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 6316–6327.

Pathiraja, B.; Liu, C.; and Senanayake, R. 2024. Fairness in Autonomous Driving: Towards Understanding Confounding Factors in Object Detection under Challenging Weather. In *Data-Driven Autonomous Driving Simulation Workshop at the IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*.

Shah, S.; Dey, D.; Lovett, C.; and Kapoor, A. 2018. *AirSim: High-Fidelity Visual and Physical Simulation for Autonomous Vehicles*, volume 5, 621–635. Cham: Springer International Publishing. ISBN 9783319673608.

Shakarian, P.; Simari, G. I.; and Bastian, N. D. 2025. Probabilistic Foundations for Metacognition via Hybrid-AI.

Vanschoren, J. 2018. Meta-Learning: A Survey. (arXiv:1810.03548). ArXiv:1810.03548.

Xi, B.; Scaria, K.; Bavikadi, D.; and Shakarian, P. 2024. Rule-Based Error Detection and Correction to Operationalize Movement Trajectory Classification. (arXiv:2308.14250). ArXiv:2308.14250.

Zhou, K.; Liu, Z.; Qiao, Y.; Xiang, T.; and Loy, C. C. 2023. Domain Generalization: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4396–4415.