Enhancing road segmentation model for asphalt edge detection

Internship final report

Bharath krishna Ramnarayan Babu
Masters Automotive technology
1820230

Supervisor & Mentors
Dr. Gijs Dubbelman
Bart Keulen - Avular Innovations

Eindhoven, Janurary 19, 2024

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Signal processing systems group
Department of Electrical Engineering

Enhancing road segmentation model for asphalt edge detection

Bharath krishna Ramnarayan Babu
1820230
Eindhoven University of technology
Eindhoven, Netherlands
b.k.ramnarayan.babu@student.tue.nl

***Abstract*—This paper investigates different out-of-distribution generalization strategies primarily using data augmentation to implement a binary segmentation model. The model aims to detect the road edge while ensuring consistent segmentation performance in different weather conditions, camera properties, and image perspectives. Out-of-distribution (OOD) generalization denotes the capability of a model to effectively and accurately predict data that deviates from the distribution encountered during its training phase. Achieving adeptness in OOD generalization is essential for ensuring the model's reliability and effectiveness in real-world deployment scenarios. Data augmentation was adopted and tested through three different experiments. The first experiment analyzes how our model architecture generalizes to OOD data during testing with the help of data augmentation. The second experiment studies the effect of various data augmentation transforms on the road edge segmentation performance. The third experiment explores the effectiveness of employing transfer learning for enhancing road edge segmentation. Our findings show that data augmentation enhances the model's performance by increasing its robustness to various operating conditions, while also promoting generalization to out-of-distribution data.**

***Index Terms*—Out-of-distribution generalization, Data augmentation**

I. Introduction

Image segmentation is one of the main computer vision problems and plays an important role in the automotive industry. In contrast to most binary segmentation tasks, where the precise segmentation of the road edge and background is not paramount, this particular project focuses on achieving high accuracy in road edge segmentation. The model is subsequently deployed on a Jetson, located on a truck, to cut asphalt at a predetermined distance from the road's edge depicted in figure 1.

Beyond accuracy, real-world decision-making systems require reliability, robustness, and safety to avoid potential mishaps. For instance, if the saw blade mounted on the truck comes too close to the edge of the road, it can break and cause damage. It is worth noting that the truck company operates a diverse fleet of trucks, which may have varying camera specifications. As such, the model must be robust to changes in camera orientation and position. Robustness against these changes described is known as out-of-distribution(OOD) generalization.

[1] [2]In deep learning, there are two types of OOD problems: near OOD, which refers to data that is somewhat similar to the training data but still lies outside the primary distribution also referred to as domain shift, and far OOD, which refers to data that is significantly different or dissimilar from the training data. It represents situations where the model encounters data that is highly unfamiliar and outside the scope of what it was trained on, also known as semantic shift. For instance, changes in the interpretation or labeling of objects within the data.

Fig. 1. Truck cutting asphalt edge

This project focuses on near OOD generalization, with a particular emphasis on a few aspects of changes in the operating environment. Specifically, changes in properties and viewpoints of the camera, and weather conditions may all be considered as instances of near OOD. This work focuses on investigating data augmentation techniques for OOD generalization [3] [4] [5].

In this work, we present an ablative study of different data augmentation techniques and show that they all enhance generalization in near OOD problems. Furthermore, we show that the augmentations have a positive impact on the IoU of our road segmentation model. We also show that adopting transfer learning strategies along with data augmentation improves the IoU of the road segmentation model further.

II. Related work

Various approaches and methodologies have been proposed in the literature to enhance the generalization of OOD samples

in machine learning models. The predominant strategy, and the main focus of this work, is data augmentation [6] [7]. Other notable approaches include transfer learning [8] [9], diverse weight averaging [10], and uncertainty modeling [11]. As discussed in the work by [12, Wei et al.], different types of image augmentation techniques are studied. The authors categorize data augmentation techniques into single-image-based augmentation and multi-image-based augmentation. Single-image-based augmentation methods are then further categorized into techniques in the areas of geometric transform, noise transform, and color transform. Multi-image-based transforms are further categorized into fusion transform and added transform. Unlike other techniques such as CutMix [13], MixUp [14], and CutOut [15], [16, Hendrycks et al.] introduces a new data augmentation technique that increases robustness to uncertainty by mixing images resulting from multiple augmentation chains. This method is important because it produces images with variety while preserving much of the image semantics and local statistics. The results of [3] show the significance of data augmentation and transfer learning strategies for OOD generalization. It has also been shown in [8] [9], that pretraining enhances the robustness of the model against OOD samples. Uncertainty modeling techniques [11] assume feature statistics to be no longer deterministic values, but probabilistic points with diverse distribution possibilities. This uncertainty in feature statistics can be leveraged to train models that are more robust against potential domain shifts, by alleviating domain perturbations. Another notable strategy is diverse weight averaging [10], It averages weights acquired from several independent training runs as models obtained from different runs are more diverse than those collected along a single run owing to differences in hyperparameters and training procedures. However, in this paper similar to the categorization presented in [12], data augmentation techniques are adopted and classified into different groups for efficient ablative experimentation.

III. Methodology

A. OOD Generalization

A commonly employed approach to evaluate the OOD generalization performance of a model is to utilize an OOD dataset during the testing phase. During the training phase, the model is trained on a specific dataset augmented with various data augmentation transforms. However, during the testing phase, a variant of the initial dataset that includes OOD elements is introduced, thereby challenging the model's ability to identify and generalize to previously unseen features. The principle behind this is that data augmentation would benefit OOD generalization by diversifying the training dataset through transformations like rotation and scaling. This variation helps the model become robust to different scenarios, improving its ability to generalize to unseen conditions and reducing overfitting. Essentially, data augmentation enhances the model's adaptability to a broader range of real-world situations. The effectiveness of the data augmentation techniques in facilitating generalization to OOD data can be quantified using generalization metrics. A detailed discussion on this topic can be found in Section IV-B1.

B. Data augmentation techniques

Three near OOD use cases are defined that need to be tackled by the data augmentation techniques they are, weather changes, variations in the position of the camera or the camera perspective, and finally, the differences in the properties of the camera itself.

To address these challenges, augmentations have been categorized into three main categories: Camera properties (color), perspective transforms (geometry), and weather changes (blur). The hypothesis is that changes in brightness, contrast, color, and so on, which are categorized as camera property changes, can help generalize the model to changes in camera properties. Similarly, Gaussian blur, noise, motion blur, and so on, which are classified as weather changes, can help generalize to changes in weather. Finally, all rotational, resizing, and perspective transformations, categorized as geometry, can tackle changes in camera perspective.(Refer fig 2)

This approach of categorizing the augmentations and mapping them to specific challenges is expected to lead to a better understanding and utilization of data augmentation techniques to improve the performance of the model in near OOD scenarios.

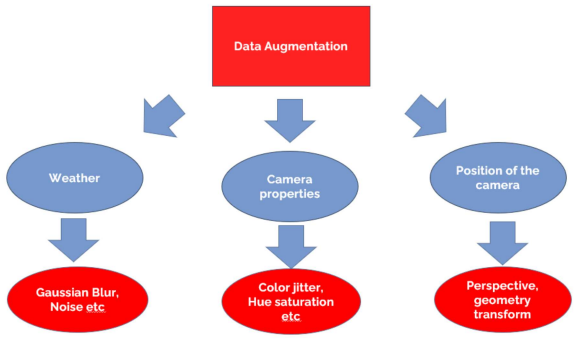

Fig. 2. Data augmentation categories

C. Transfer learning

The second approach being investigated is transfer learning. As elucidated in section II, research has also established the benefits of incorporating transfer learning methodologies, particularly in improving OOD generalization performance. Thus, another approach is to pre-train the model on a larger road segmentation dataset and subsequently utilize the learned weights for fine-tuning on the project-specific dataset, which is comparatively smaller in size. This technique ensures effective knowledge transfer, significantly improving both the segmentation and the generalization performance of the model.

IV. Experiments

A. Experimental setup

*1) **Architecture**:* The standard vanilla UNet architecture [17] was used as the segmentation head and the ResNet

TABLE I
HYPERPARAMETERS USED

Hyperparameters	Value
Optimizer	Adam
Learning rate	0.0001
Epoch	30
Batch size	4
Activation function	Sigmoid

encoder [18] pre-trained on the ImageNet dataset as the backbone [19], considering there is the need to keep the architecture simple and lightweight for two reasons: first, to deploy on jetson real-time and second, to reduce the number of parameters. The hyperparameters used were kept constant throughout all experiments to study the differences in performance arising from different augmentations while keeping the hyperparameters constant. Refer I for the hyperparameters used.

*2) **Datasets**:* The following four datasets were used in the experiments:

- The **CAMVID** dataset [20]is a prominent dataset used for road segmentation tasks comprising over 700 images with dimensions of 320 pixels height and 480 pixels width.
- The **CAMVID OOD** dataset [21] contains the same number of images as the original CamVid test dataset. However, each image contains an out-of-distribution animal.
- The **Cityscapes** dataset [22]is a large dataset with over 5000 finely annotated images that primarily serves as a benchmark for road segmentation tasks.
- The **ABX truck dataset** is a custom dataset created for the project with over 1100 images in diverse environmental conditions taken from a truck to perform various computer vision tasks.

Multiple datasets are used to verify if the findings are robust against different settings, rather than being limited to a single dataset. Detailed information on the dataset split used is given under appendix VIII.

*3) **Metrics and loss functions**:* To study the generalization performance we use three standard metrics of OOD generalization: area under the ROC curve (AUROC), the false positive rate at 95% recall (FPR95), and area under the precision-recall curve (AUPR) [23] [24] [25] [26].

1) **AUROC** score: It measures the ability of a model to distinguish between positive and negative samples, and is often used to evaluate the performance of models in detecting OOD samples.
2) **AUPR** score: AUPR is often considered more informative than AUROC, especially when dealing with imbalanced datasets.
3) **FPR@95% TPR:** It indicates how well the model can maintain a high rate of true positives while keeping false positives under control.

For the evaluation of the segmentation performance, IoU was used as the primary metric. While IoU is a commonly used metric for evaluating image segmentation performance, it may not fully account for intensity and structural changes between the prediction and ground truth masks. Hence, the Structural Similarity Index (SSIM) is used to quantify the similarity between two images. F1 score, F2 score, precision, and recall were used as secondary metrics. After testing the model with several loss functions, we used Dice Loss as it had the best tradeoff between the interpretability of the results and the performance of the model [27] [28]. Twersky loss and Lovasz loss were also considered for experiment 1.

B. Experiments

*1) **OOD Generalization**:* The initial strategy III-A was adopted for the first experiment. The road class within the CAMVID dataset was utilized as the target for training a binary segmentation model, and the CAMVID OOD dataset for testing. The CAMVID dataset was trained as a binary segmentation task to detect the road and the background with different data augmentations applied and later tested on CAMVID OOD to study the OOD robustness of the architecture. AUROC, AUPR, and FPR@95% TPR were used to assess the OOD generalizability of the model. The findings of this study were aimed at observing the model's performance drop when confronted with an OOD element and its generalization capability during the testing phase.

*2) **Data augmentation**:* The CAMVID dataset provides a sound basis for testing the OOD generalization capability. Nonetheless, given its restricted size, it is imperative to reinforce these conclusions with supplementary testing. We use the Cityscapes dataset for further analysis and evaluation. In the second experiment, we analyzed the impact of three distinct categories of data augmentation techniques, namely color, geometry, and blur, as outlined in section III-B. Therefore, the Cityscapes dataset was trained as a binary segmentation task to detect the road and the background with augmentations applied individually and in all possible combinations. This analysis aimed to evaluate the effectiveness of each group of augmentations in enhancing the quality of the dataset.

The ABX dataset is used as the test dataset. In other words, the model is directly evaluated on the ABX dataset, which comprises images with unknown features, such as grass fields and canals that are not present in the Cityscapes dataset. Additionally, the efficacy of various data augmentation techniques is also analyzed to determine the most effective combination. SSIM is employed as a complementary qualitative metric alongside IoU to assess image quality. This experiment was run for 50 epochs while the other hyperparameters remain the same as in I.

*3) **Fine tuning**:* The model with the highest IoU in Experiment 2 as discussed in IV-B2, is then used to fine-tune on the ABX dataset to see if it improves the segmentation performance on the ABX dataset when compared to the training of the ABX dataset from scratch. The model trained without augmentation on Cityscapes, and the ABX dataset trained from scratch with and without augmentation are also evaluated for comparison (Refer II). The experiment is conducted for

TABLE II
MODELS FOR COMPARISON

Model name	Description
FT_aug	Model pre-trained on Cityscapes with augmentations and then fine-tuned on ABX
FT_noaug	Model pre-trained on Cityscapes without augmentations and then fine-tuned on ABX
ABX_aug	Model trained on ABX directly with augmentations
ABX_noaug	Model trained on ABX directly without augmentations

different training dataset sizes of 50, 100, 250, 500, and 750 images respectively. This is done to see how easily the learned features can be adapted to the new settings. It is also beneficial when the environment changes, for instance, the installation of a new camera. The model has to be fine-tuned for this new situation with as little labeled data as possible. The test and validation sizes remain constant, and a random subset of the remaining images is used for training.

V. RESULTS

A. *OOD generalization*

The results of the OOD generalization experiment as explained in IV-B1 of the ResNet encoder with UNet architecture are shown (Table III). Upon testing the model with OOD elements, it was observed that the false positive rate (FPR) value increased by a negligible amount. This indicates that the model's capability to maintain low false detections is good. Figure 3 depicts an example image of the model's ability to generalize to OOD objects. Additional information regarding the comparison of various augmentation frameworks and loss metrics can be found in appendix VII.

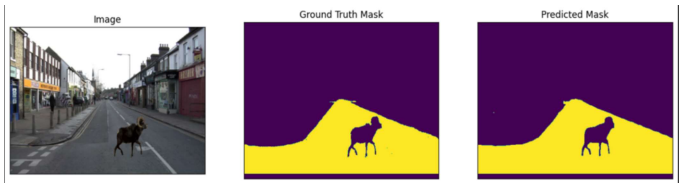

Fig. 3. OOD generalization result: Ground truth mask(left), Predicted mask(right)

B. *Data Augmentation*

Fig. 4. Model trained with Cityscapes(No augmentation) IV and tested on ABX dataset. From left: Ground truth mask(4b) and the predicted mask(4c)

The results presented in Table IV indicate a significant improvement in the model performance when augmentation is applied on Cityscapes and tested on ABX. The IoU increases

Fig. 5. Model trained with Cityscapes(All augmentation IV) and tested on ABX dataset. From left: Ground truth mask(5b) and the predicted mask(5c)

Fig. 6. Model trained with Cityscapes(No augmentation) IV and tested on ABX dataset. From left: Ground truth mask(6b) and the predicted mask(6c)

Fig. 7. Model trained with Cityscapes(All augmentation IV) and tested on ABX dataset. From left: Ground truth mask(7b) and the predicted mask(7c)

from 67 to 74.6. The visual comparison of the samples presented in Figure 4 and 5 further supports this observation. However, it is noteworthy that the SSIM values do not show a major difference between the different augmentations, which could be attributed to class imbalance. The overlap of the mask and the original image reveals that the background constitutes the major part of the image, hence resulting in minimal difference.

The interesting aspect to note here is that the model tries to generalize to the ABX dataset using the features learned from the model trained on Cityscapes. For instance, in Figure 6, the model classifies both the main road and side road as asphalt roads, whereas the ground truth is only the main road. This suggests that the model is relying on the learned features from the Cityscapes dataset, which might not be sufficient for the ABX dataset. In contrast, Figure 7 demonstrates that applying data augmentation to the Cityscapes dataset enables the model to learn the ground truth, i.e., only straight roads.

C. *Fine tuning*

Figure 8 shows the graph for the IoU of all four models mentioned in section IV-B3. It is observed that training on ABX without fine-tuning is already providing satisfactory results for this particular use case when working with a small number of images(IoU of 93.5 with 50 images). The reason for this might be the fact that a pre-trained ResNet50 encoder is used, which seems to be capable enough already. Furthermore, another conclusion is that fine-tuning on the total number of images provides marginally better results (IoU of 95.7).

TABLE III
OOD GENERALIZATION RESULTS ON CAMVID DATASET

Model	Encoder	Augmentation framework	Loss	IoU (In/OOD)	AUROC (%) (In/OOD)	AUPR (%) (In/OOD)	FPR @95% TPR (%) (In/OOD)
UNet	ResNet18	Albumentations [29]	Dice	90.7/87.4	98/97	95/93.7	3/3.86

TABLE IV
DIFFERENT DATA AUGMENTATION RESULTS TESTED ON ABX

Augmentation	IoU (Test on ABX)	Average SSIM (Test)
None	67	81.2
Color	69	83
Blur	68.4	83.9
Geometry	64	83.4
Color and Blur	70.9	84.8
Color and Geometry	72.9	84.2
Blur and Geometry	73.7	85.6
All	**74.6**	**85.7**

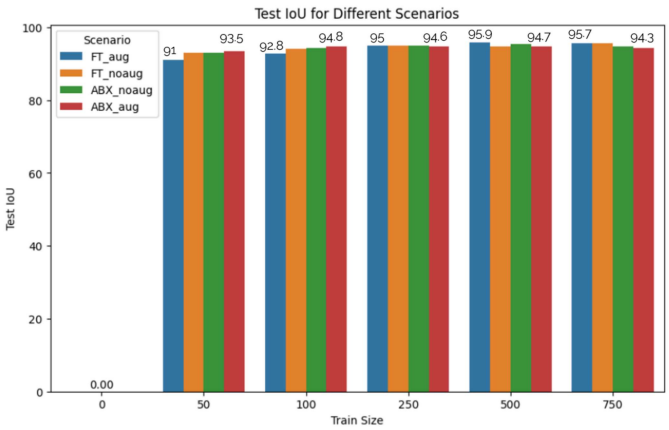

Fig. 8. Test IoU for varying number of images for all the 4 models mentioned in section IV-B3

VI. CONCLUSION

This study shows that using data augmentation techniques improves the performance and robustness of a binary segmentation model. Results on the CAMVID OOD dataset (section V-A) demonstrate the model's ability to adapt to near OOD cases. Different data augmentation combinations effectively improve segmentation performance under various conditions. Visualizing intermediate layers of the model can provide further insights into the reasons behind these observations. Based on the results from V-C, it can be inferred that open-source segmentation datasets hold the potential to be employed for direct evaluation on custom datasets, as long as they pertain to the specific task and exhibit similarity in training distribution. Pre-training on such open source datasets could also improve the performance through transfer learning and can be more efficient depending on the similarity of the dataset distribution. Even though there are synthetic datasets such as Rain Cityscapes [30] and Foggy Cityscapes [31] for different weather conditions, a shortcoming is that there are no realistic datasets available to evaluate data augmentation techniques for individual augmentation categories such as weather, camera properties, and position of the camera (Refer fig 2).

A. Future work

There are several aspects for further research in which the performance of the model can be improved. One method is to develop a new benchmark/dataset with images in various conditions such as a mixture of Wildash, Cityscapes, Berkeley deep drive, etc. A rich and diverse dataset ensures that many features are learned, resulting in better OOD generalization. One another approach is to use a combination of losses to improve the results further.

REFERENCES

[1] J. Zhang, J. Yang, P. Wang, H. Wang, Y. Lin, H. Zhang, Y. Sun, X. Du, K. Zhou, W. Zhang *et al.*, "Openood v1. 5: Enhanced benchmark for out-of-distribution detection," *arXiv preprint arXiv:2306.09301*, 2023.

[2] J. Yang, P. Wang, D. Zou, Z. Zhou, K. Ding, W. Peng, H. Wang, G. Chen, B. Li, Y. Sun *et al.*, "Openood: Benchmarking generalized out-of-distribution detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32 598–32 611, 2022.

[3] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo *et al.*, "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8340–8349.

[4] A. Mumuni and F. Mumuni, "Data augmentation: A comprehensive survey of modern approaches," *Array*, p. 100258, 2022.

[5] F. Wenzel, A. Dittadi, P. Gehler, C.-J. Simon-Gabriel, M. Horn, D. Zietlow, D. Kernert, C. Russell, T. Brox, B. Schiele *et al.*, "Assaying out-of-distribution generalization in transfer learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 7181–7198, 2022.

[6] V. Olsson, W. Tranheden, J. Pinto, and L. Svensson, "Classmix: Segmentation-based data augmentation for semi-supervised learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1369–1378.

[7] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T.-Y. Lin, E. D. Cubuk, Q. V. Le, and B. Zoph, "Simple copy-paste is a strong data augmentation method for instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2918–2928.

[8] D. Hendrycks, K. Lee, and M. Mazeika, "Using pre-training can improve model robustness and uncertainty," in *International conference on machine learning*. PMLR, 2019, pp. 2712–2721.

[9] A. E. Orhan, "Robustness properties of facebook's resnext wsl models," *arXiv preprint arXiv:1907.07640*, 2019.

[10] A. Rame, M. Kirchmeyer, T. Rahier, A. Rakotomamonjy, P. Gallinari, and M. Cord, "Diverse weight averaging for out-of-distribution generalization," *Advances in Neural Information Processing Systems*, vol. 35, pp. 10 821–10 836, 2022.

[11] X. Li, Y. Dai, Y. Ge, J. Liu, Y. Shan, and L.-Y. Duan, "Uncertainty modeling for out-of-distribution generalization," *arXiv preprint arXiv:2202.03958*, 2022.

[12] J. Wei, Q. Wang, X. Song, and Z. Zhao, "The status and challenges of image data augmentation algorithms," in *Journal of Physics: Conference Series*, vol. 2456, no. 1. IOP Publishing, 2023, p. 012041.

[13] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[14] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[15] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[16] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "Augmix: A simple data processing method to improve robustness and uncertainty," *arXiv preprint arXiv:1912.02781*, 2019.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[19] Popular segmentation models. [Online]. Available: https://paperswithcode.com/task/semantic-segmentation

[20] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in *ECCV (1)*, 2008, pp. 44–57.

[21] V. Besnier, A. Bursuc, D. Picard, and A. Briot, "Triggering failures: Out-of-distribution detection by learning from local adversarial attacks in semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 701–15 710.

[22] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[23] D. Hendrycks, S. Basart, M. Mazeika, A. Zou, J. Kwon, M. Mostajabi, J. Steinhardt, and D. Song, "Scaling out-of-distribution detection for real-world settings," *arXiv preprint arXiv:1911.11132*, 2019.

[24] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*, 2016.

[25] J. Chen, Y. Li, X. Wu, Y. Liang, and S. Jha, "Robust out-of-distribution detection for neural networks," *arXiv preprint arXiv:2003.09711*, 2020.

[26] J. Liu, Z. Shen, Y. He, X. Zhang, R. Xu, H. Yu, and P. Cui, "Towards out-of-distribution generalization: A survey," *arXiv preprint arXiv:2108.13624*, 2021.

[27] S. Jadon, "A survey of loss functions for semantic segmentation," in *2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB)*. IEEE, 2020, pp. 1–7.

[28] B. Liu, J. Dolz, A. Galdran, R. Kobbi, and I. B. Ayed, "Do we really need dice? the hidden region-size biases of segmentation losses," *Medical Image Analysis*, vol. 91, p. 103015, 2024.

[29] E. K. V. I. I. A. Buslaev, A. Parinov and A. A. Kalinin, "Albumentations: fast and flexible image augmentations," *ArXiv e-prints*, 2018.

[30] X. Hu, C.-W. Fu, L. Zhu, and P.-A. Heng, "Depth-attentional features for single-image rain removal," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2019, pp. 8022–8031.

[31] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *International Journal of Computer Vision*, vol. 126, pp. 973–992, 2018.

VII. Appendix A

TABLE V
COMPARISON OF GENERALIZATION RESULTS USING DIFFERENT AUGMENTATION FRAMEWORKS

Model	Encoder	Augmentation framework	Loss metric	IoU (IN/OOD)	AUROC (%) (IN/OOD)	AUPR(%) (IN/OOD)	FPR (%) (@95% TPR)
UNet	ResNet18	Torchvision transforms	Dice	91/89.1	98/97	95.6/94.7	2.2/2.35
			Twersky	90.5/88	98/97	95.3/94.1	2.63/2.93
			Lovasz	**92.6/90.5**	**97/97**	**96.6/95.6**	**1.21/1.23**

Model	Encoder	Augmentation framework	Loss metric	IoU (IN/OOD)	AUROC (%) (IN/OOD)	AUPR(%) (IN/OOD)	FPR (%) (@95% TPR)
UNet	ResNet18	Albumentations	Dice	90.7/87.4	98/97	95/93.7	3/3.86
			Twersky	90.4/86.8	97/97	95/93	3.02/3.86
			Lovasz	**92.2/89.4**	**97/96**	**96.6/95.4**	**0.98/1.07**

The experimentation conducted using two data augmentation frameworks revealed that the Lovasz loss metric produced a higher Intersection over Union (IoU) than both Twersky and Dice loss metrics. Although the augmentation frameworks showed minimal performance differences, the torchvision transforms exhibited marginally better results in most cases. While the same augmentation transforms were applied using both frameworks, the probability of their application to images might have been different for some operations, leading to the observed difference in performance.

A notable point is that the augmentations applied were kept constant for all the experiments performed. However, the experiment's results only applied to training with the CAMVID dataset, which consisted of only 700 images. Moreover, as subsequent experiments did not use this dataset, the findings were not coherent with the other experiments performed. Therefore, the focus of this experiment was on the model architecture's inherent generalization capability.

VIII. Appendix B

TABLE VI
DATASET SPLIT FOR TRAINING

Experiments	**Datasets used**	**Dataset split**		
		Train	Val	Test
OOD generalization	CAMVID for training and CAMVID OOD for testing	370	100	230
Data augmentation	Cityscapes for training and ABX for testing	3000	1000	1100
Fine tuning	ABX	[0, 50, 100, 250, 500, 750]	130	130