PAPER 04
Camera- and LiDAR-based Weather-type Classification for Autonomous Vehicles

Siddhant Som[1], Jesper Holmblad[1], and Eren Erdal Aksoy[1]

***Abstract*—Autonomous vehicles (AV) must have the ability to perceive and comprehend their surroundings accurately to ensure safe operation. They rely on various sensors to perceive their environment and make logical decisions. Detecting weather conditions in the surrounding environment is a crucial task for AVs to achieve optimal control. This paper demonstrates the impact of combining data from two different modalities, RGB Camera and LiDAR, on the performance of a vision transformer, a popular neural network for working with images, for the task of weather-type detection. To establish baselines, we also perform experiments using only RGB data, followed by only LiDAR counterpart. The experiments are conducted using a part of the Zenseact Open Dataset (ZOD) dataset.**

***Index Terms*—Weather-type detection, LiDAR, RGB**

I. Introduction

Modern autonomous vehicles excel in optimal weather conditions, yet their performance often falters in adverse weather scenarios, such as heavy rain, fog, and snow, where reduced visibility poses significant safety challenges. In response, researchers have been exploring novel approaches to equip these vehicles with robust adaptive capabilities tailored to diverse environmental conditions.

Recent advancements in deep learning techniques have spurred the development of sophisticated weather classification systems. For instance, primitive networks with minimal training time have been employed for cloudy/clear weather classification [1]. Furthermore, the integration of ambient light and color analysis into a Faster R-CNN framework has shown promise in enhancing vehicle detection under varying weather conditions [2]. Additionally, efforts have been made to leverage neural networks and k-Nearest Neighbors algorithms for weather prediction and classification, albeit with varying degrees of complexity [3].

There have been many advancements in the field of image processing, however, the majority of the State Of The Art (SOTA) in weather perception does not leverage these advancements [1] [2] [3] [4]. One notable development is the use of attention in models such as vision transformers [5]. Exploring the applicability of such techniques seems worthwhile. Aside from accuracy, another aspect of AV is robustness. Some solutions make decisions based on single modality [4]. This approach is more vulnerable to malfunction than a multi-modal one. There are papers based on combining multiple simplistic sensors such as temperature and humidity [3]. We recognize the cost efficiency of such solutions but aim to provide a higher-performance solution. An older study claims that the combination of LiDAR and the camera is "well suited" [6].

Acknowledging the growing significance of multimodal datasets in autonomous driving research, recent works have emphasized the importance of comprehensive datasets that encompass various weather and lighting conditions. For instance, the Zenseact Open Dataset (ZOD) [7] offers a diverse range of real-world scenarios collected from European regions, facilitating a more nuanced understanding and modeling of adverse weather conditions. Moreover, advancements in sensor fusion, particularly the integration of camera and LiDAR data, have shown promise in enhancing weather recognition and overall system robustness [6].

Inspired by these developments, we propose a multi-modal weather classifier tailored specifically for autonomous driving applications. By integrating insights from diverse sources and leveraging state-of-the-art techniques, our method aims to enhance the adaptability and safety of autonomous vehicles in adverse weather conditions such as rain, fog, and snow.

II. Related Work

This section mentions works that have proposed datasets or detection networks for adverse weather conditions.

Adverse weather conditions such as the presence of heavy rain, fog, and snow reduce visibility and may affect driving safety. In recent times, multimodal datasets have become increasingly popular as Autonomous Driving (AD) systems aim to use data from multiple sensors to understand their surroundings more accurately. The Waymo open [8] multimodal dataset consists of annotated LiDAR, as well as camera data. It has 1150 sequences spanning 20 seconds each and covers a comprehensive range of driving scenarios, with huge geographical diversity. It is, however, limited by the types of weather conditions covered in the data, as it only consists of scenarios that have clear weather or rain. It also lacks pointwise annotations but has 3D bounding boxes for all annotated objects. KITTI [9], released in 2012 is regarded as one of the most impactful multimodal AD datasets. It consists of 6 hours of traffic scenarios, collected using stereo cameras, LiDAR and high precision inertial navigation system. It is limited by its diversity and the types of weather conditions covered. However, with 200K object labels in the form of 3D tracklets, KITTI enabled advancements in AD research in

[1]Halmstad University, School of Information Technology, Center for Applied Intelligent Systems Research, Halmstad, Sweden

This work was funded by the European Union (grant no. 101069576).Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them.

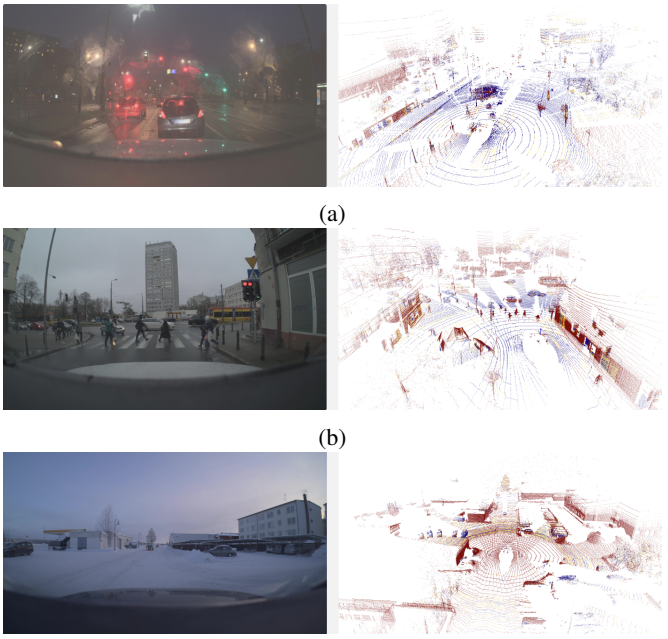
areas such as 3D object detection and tracking, as well as visual odometry. Using KITTI as a basis, the SemanticKITTI [10] was released in 2019 by Behley et al., which is very useful for the task of semantic segmentation. It has point-wise annotation with 28 classes, achieved by annotating all 22 sequences of the KITTI Vision Odometry Benchmark [11], consisting of over 43000 scans. SemanticKITTI is limited by its modality and diversity in weather conditions as it only consists of LiDAR data and covers only clear weather scenarios. NuScenes [12], released in 2019 is one of the first publicly available datasets and consists of 3 modalities - camera, LiDAR and Radio Detection And Ranging (RADAR), with a full 360-degree coverage. It consists of 1000 scenes, each 20 sec long and fully annotated with 3D bounding boxes for 23 classes and 8 attributes. It has 7× times the annotations and 100× times the images of the KITTI dataset. It also covers rainy, snowy, and foggy weather conditions. ZOD [7] is a relatively new large-scale, diverse, and multimodal dataset that has data collected from several European countries. It consists of data collected using high-resolution sensors covering a wide range of traffic scenarios, weather conditions, road types, and lighting conditions. It comprises Frames, Sequences, and Drives and is explained in detail in the section III.

When it comes to the detection architectures, the Vision transformer introduced by Dosovitskiy et al. [5] attains excellent results compared to state-of-the-art Convolutional Neural Networks (CNN) while requiring few computational resources for training. Dannheim et al. [6] show how the combination of data from the camera and LiDAR helps improve weather recognition in AD systems. This work, however, relies on conventional feature extraction methods without applying advanced neural networks. Zhong et al. [13] provide a comprehensive summary of popular fusion methods used to combine camera and LiDAR data, while Vargas et al. [14] discuss various autonomous sensors including camera and LiDAR, discussing their advantages, disadvantages, and robustness to adverse weather conditions.

III. Dataset

To carry out the experiments in this project, the multimodal ZOD [7] was used. The dataset comprises a large set of Frames, Sequences, and Drives. Frames consist of 100,000 curated camera images, Sequences consist of 1473 20-second videos and Drives consist of 29 videos, each of which is a few minutes long. Each subset consists of high-quality camera data as well as accompanying LiDAR data. A subset of 10,000 Frames were used in the project. The data in ZOD is collected from 14 European countries over multiple years and is curated to contain several traffic scenarios, weather conditions, road types, and lighting conditions. Fig.1 shows sample camera and LiDAR data for different weather types.

The ZOD Frames are fully annotated with the Object, Lane, Road Condition, and Weather Type information, as well as semantic segmentation masks and 2D/3D bounding boxes. There exist five different Weather Type annotations in ZOD: clear, cloudy, rain, fog, and snow. The distribution of these



(a) (b) (c)

Fig. 1: Sample RGB camera images (left) and corresponding LiDAR point clouds (right) for rainy, cloudy, and snow weather types, respectively.

labels is depicted in Fig. 2. This figure conveys the fact that there is a certain imbalance in the amount of annotated data for each weather type. Despite of this imbalance, the diversity in terms of types of weather, time of day, and locations made ZOD a solid choice to conduct our experiments, as opposed to other datasets such as nuScenes [12] and Waymo [8].

IV. Approach

Since the original dataset does not have an even distribution of samples concerning the weather type labels, a subset of the dataset was chosen. In this new ZOD subset, the distribution of samples of all weather types is roughly equal. More specifically, this subset has the following number of images in each weather class: 'cloudy': 2000, 'clear': 2000, 'snow':

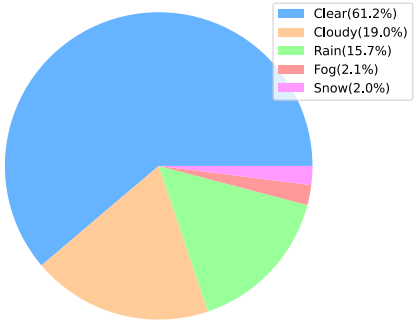


Fig. 2: Distribution of Weather type labels in ZOD [7].

1972, 'rain': 2000, and 'fog': 2000. We employed 20% of these images in the test set. Experiments were conducted for the three following scenarios:

1) Only RGB data
2) Only LiDAR data
3) A combination of RGB and LiDAR data

.

The model employed in our experimental evaluation is a pre-trained vision transformer [5] from the package 'PyTorch Image Models' or 'timm'. In each scenario, the model was trained for 50 epochs using the Adagrad optimizer with a learning rate of 0.009. The loss function used for training the model is the cross entropy loss function.

A. RGB Only

The RGB only (i.e., camera data) in the dataset was captured by an 8MP high-resolution camera. The original size of each image is $3 \times 3848 \times 2168$. Fig. 3 shows some sample RGB camera images. Before being passed to the vision transformer, the image is resampled to $3 \times 224 \times 224$ pixels due to the pre-trained vision transformer restrictions.

B. LiDAR Only

The LiDAR data is captured by 3 LiDAR sensors - 2×Velodyne VLP16 and 1×Velodyne VLS128 - and each point cloud is represented by a 6-dimensional vector with the timestamp, 3D coordinates (x, y, z), intensity value, and diode index. Each LiDAR point cloud contains 254K points on average.

Following the work in [15], we converted the raw LiDAR point cloud into a Panoramic View (PV) image representation. More specifically, the 3D points (x,y,z) are projected onto a spherical surface and the range value is computed. In the projected form, we represent each point cloud as a 5-dimensional point cloud with the 3D coordinates (x,y,z), range, and intensity. The Horizontal Field of View (HFOV) of the LiDAR sensors used is 360°. This panoramic image representation of a LiDAR point cloud covers the full 360°. The dimensions of the PV of a LiDAR point cloud after projection is $5 \times 128 \times 2048$. Fig. 4 shows a sample projection, i.e., a PV, of a LiDAR point cloud.

In addition to the PV, we also cropped the Front View (FV) of a LiDAR point cloud, covering only the middle 120°. The dimensions of the FV of a LiDAR point cloud after projection is $5 \times 128 \times 683$. We conduct experiments with both the PV and the FV of the LiDAR point clouds. In both cases, the dimensions of the input expected by the vision transformer are $5 \times 224 \times 224$.

Fig. 3: Sample RGB data.

The point clouds are resized in two ways: either through enhancement or compression. In the enhancement mode, a Convolutional Neural Network (CNN) processes the input and reshapes it to $5 \times 224 \times 224$ after a set of convolution operations. It ensures that essential information is captured before the input is reshaped. The context module from SalsaNext [15] is also combined with the enhancer to capture rich contextual information. This context module uses a dilated convolution stack that fuses a larger receptive field with a smaller one by adding 1x1 and 3x3 kernels at the beginning. This captures global context alongside more detailed spatial information. The compression mode simply reshapes the input to the desired size directly. Thus, we have in total 4 experimental protocols to follow, when working with LiDAR data:

1) PV+Enhancement
2) PV+Compression
3) FV+Enhancement
4) FV+Compression

C. RGB+LiDAR

Once we conduct individual experiments with RGB and LiDAR data, we go a step further and fuse both modalities. In this final scenario, we follow an early fusion scheme, i.e., the 3 channels of the camera image data are concatenated with the 5 channels of a corresponding LiDAR point cloud, resulting in 8 channels. Thus, the input dimensions to the vision transformer become $8 \times 224 \times 224$. Fig. 5 shows the flow diagram of how RGB and LiDAR data are combined. In this fused version, there exist 4 possible experimental protocols:

1) RGB + (PV+Enhancement)
2) RGB + (PV+Compression)
3) RGB + (FV+Enhancement)
4) RGB + (FV+Compression)

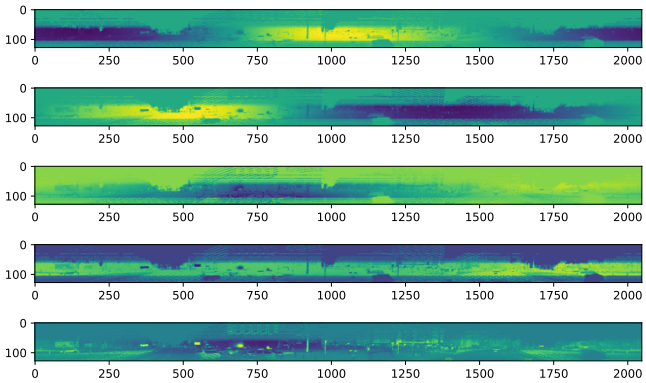


Fig. 4: Panoramic View (i.e., spherical projection) of a LiDAR point cloud. Each row represents one unique channel: 3D coordinates (x,y,z), range, and intensity, respectively.

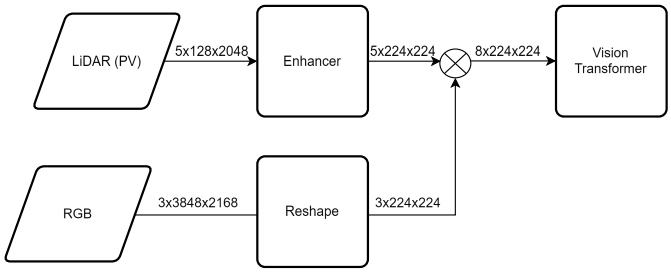


Fig. 5: Fusion of RGB and LiDAR data.

In the scenarios involving the data enhancement, a 2-stage training method was used. In the first stage, after combining the projected LiDAR and the corresponding RGB data all values in the 3 RGB channels are set to 0, and then the model consisting of both the enhancer and the vision transformer is trained. In the second stage, the parameters of the enhancer are frozen and the training is repeated with the values in the RGB channel reset to their original values. Now, only the parameters of the vision transformer are updated. This 2-stage training is employed because the enhancer plays no role in processing the RGB data, and hence its involvement in backpropagation using both RGB and LiDAR data might be detrimental to its training.

V. Results

The metric used for evaluation is accuracy. This is treated as a multi-class problem with 5 labels. The test set, consisting of 1974 images - which is 20% of the dataset - is employed to evaluate the model in each scenario. The accuracy is computed as the ratio between the number of correctly predicted labels - with 5 possible values - and the total number of images in the test set. Obtained results are provided in Table I.

a) RGB only: Using only RGB data, the accuracy on the test set achieved was 81.14% (See the first row in Table I). Based on the confusion matrix shown in Fig. 6a, we see that there is some ambiguity between the snow and rain classes. However, the model predicts all classes with reasonable accuracy with limited training, with the best ones being clear and fog.

b) LiDAR only: Each LiDAR data file can be transformed either into a PV or a FV, using a projection function. The dimensions of the PV are 5x128x2048 while that of the FV are 5x128x682. Each file is then reshaped to have dimensions 5x224x224 before being fed to the model. The FV width is computed as the middle one-third section of PV width. Recall the fact that there are 4 subcases when using only LiDAR data: PV+Compression, PV+Enhancement, FV+Compression, and FV+Enhancement. As shown between rows 2 and 5 in Table I, the best test accuracy score (78.56%) was obtained in the case PV+Enhancement. This accuracy score is slightly lower than that of RGB only approach. Fig. 6b depicts the corresponding confusion matrix.

In all cases, the model does extremely well in predicting the fog class. However, the model struggled with the rain class and was much worse at predicting it when compared to the case of using RGB data.

TABLE I: Results - all modalities

Modality	Train Acc	Test Acc
RGB	93.85	81.14
LiDAR - PV + Compression	97.21	74.30
LiDAR - PV + Enhancement	87.84	78.56
LiDAR - FV + Compression	89.43	75.16
LiDAR - FV + Enhancement	85.41	73.39
LiDAR - PV + Compression+RGB	**87.74**	**81.39**
LiDAR - PV + Enhancement+RGB	84.92	71.51
LiDAR - FV + Compression+RGB	76.14	69.39
LiDAR - FV + Enhancement+RGB	93.47	75.2

c) RGB+LiDAR: When it comes to the early fusion of RGB and LiDAR modalities, the best result (81.39%) was obtained for the case LiDAR - PV + Compression+RGB as shown in row 6 in Table I. The obtained confusion matrix is depicted in Fig. 6c.

The model performs reasonably across all classes, however, there it experiences some ambiguity when predicting snow. The performance on the rain class improved due to the presence of the RGB modality. The results of all modalities given in Table I convey the fact that the improvement in the fused model is marginal in contrast to the case of using the RGB-only modality.

VI. Conclusion

AVs rely on multiple sensors to perceive their surroundings and make logical decisions. This is essential to ensure safety on roads. In this paper, we have demonstrated how fusing data from 2 modalities - RGB and LiDAR can influence a vision transformer for the task of weather detection for 5 different kinds of weather. We used data from the ZOD dataset. We compare the results in multiple cases - using only RGB data, using only LiDAR data, and then finally fusing RGB and LiDAR data. In the cases involving only LiDAR data, we further investigated the differences in results when using PV and FV, as well as compression and enhancement. Our experiments showed that a fusion of the PV of LiDAR, subjected to compression and RGB data yielded the best accuracy in detecting weather. The contribution of the LiDAR modality in our early fusion scheme is marginal.

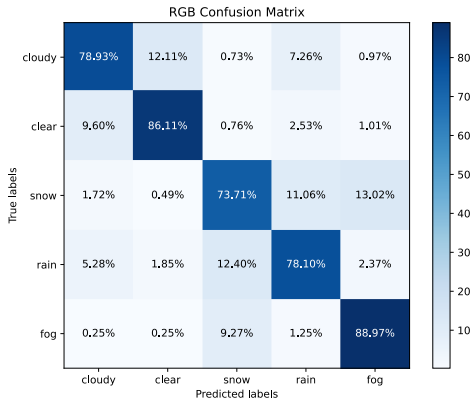
References

[1] M. Kalkan, G. E. Bostancı, M. S. Güzel, B. Kalkan, Ş. Özsarı, Ö. Soysal, and G. Köse, "Cloudy/clear weather classification using deep learning techniques with cloud images," *Computers and Electrical Engineering*, vol. 102, p. 108271, 2022.

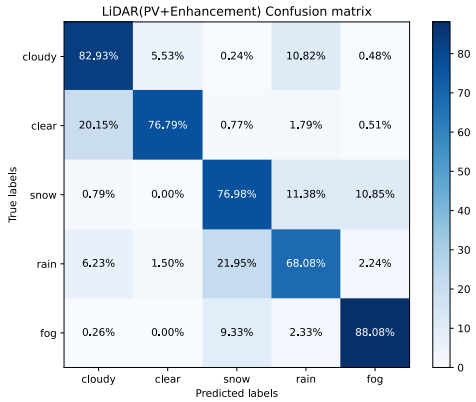
[2] E. Tian and J. Kim, "Improved vehicle detection using weather classification and faster r-cnn with dark channel prior," *Electronics*, vol. 12, no. 14, p. 3022, 2023.

[3] R. Mantri, K. R. Raghavendra, H. Puri, J. Chaudhary, and K. Bingi, "Weather prediction and classification using neural networks and k-nearest neighbors," in *2021 8th International Conference on Smart Computing and Communications (ICSCC)*. IEEE, 2021, pp. 263–268.

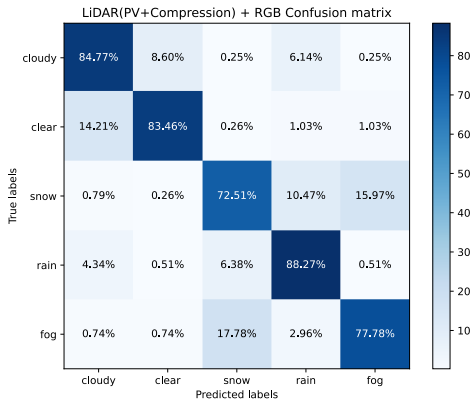
[4] M. M. Dhananjaya, V. R. Kumar, and S. Yogamani, "Weather and light level classification for autonomous driving: Dataset, baseline and active learning," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021, pp. 2816–2821.



(a)



(b)



(c)

Fig. 6: Confusion matrices for the modalities (a) RGB only (b) LiDAR (PV + Enhancement) (c) LiDAR (PV + Enhancement)+RGB

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[6] C. Dannheim, C. Icking, M. Mäder, and P. Sallis, "Weather detection in vehicles by means of camera and lidar systems," in *2014 Sixth International Conference on Computational Intelligence, Communication Systems and Networks*, 2014, pp. 186–191.

[7] M. Alibeigi, W. Ljungbergh, A. Tonderski, G. Hess, A. Lilja, C. Lindstrom, D. Motorniuk, J. Fu, J. Widahl, and C. Petersson, "Zenseact open dataset: A large-scale and diverse multimodal dataset for autonomous driving," 2023.

[8] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, S. Zhao, S. Cheng, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov, "Scalability in perception for autonomous driving: Waymo open dataset," 2020.

[9] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[10] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "Semantickitti: A dataset for semantic scene understanding of lidar sequences," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9297–9307.

[11] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[12] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[13] H. Zhong, H. Wang, Z. Wu, C. Zhang, Y. Zheng, and T. Tang, "A survey of lidar and camera fusion enhancement," *Procedia Computer Science*, vol. 183, pp. 579–588, 2021, proceedings of the 10th International Conference of Information and Communication Technology. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050921005767

[14] J. Vargas, S. Alsweiss, O. Toker, R. Razdan, and J. Santos, "An overview of autonomous vehicles sensors and their vulnerability to weather conditions," *Sensors*, vol. 21, no. 16, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/16/5397

[15] T. Cortinhal, G. Tzelepis, and E. Erdal Aksoy, "Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds," in *Advances in Visual Computing: 15th International Symposium, ISVC 2020, San Diego, CA, USA, October 5–7, 2020, Proceedings, Part II 15*. Springer, 2020, pp. 207–222.