Channel Boosting Feature Ensemble for Radar-based Object Detection

Shoaib Azam, Farzeen Munir and Moongu Jeon
School of Electrical Engineering and Computer Science
Gwangju Institute of Science and Technology Gwangju, South Korea
Email: (shoaibazam, farzeen.munir, mgjeon)@gist.ac.kr

Abstract—**Autonomous vehicles are conceived to provide safe and secure services by validating the safety standards as indicated by SOTIF-ISO/PAS-21448 (Safety of the intended functionality)[1]. Keeping in this context, the perception of the environment plays an instrumental role in conjunction with localization, planning and control modules. As a pivotal algorithm in the perception stack, object detection provides extensive insights into the autonomous vehicle's surroundings. Camera and Lidar are extensively utilized for object detection among different sensor modalities, but these exteroceptive sensors have limitations in resolution and adverse weather conditions. In this work, radar-based object detection is explored provides a counterpart sensor modality to be deployed and used in adverse weather conditions. The radar gives complex data; for this purpose, a channel boosting feature ensemble method with transformer encoder-decoder network is proposed. The object detection task using radar is formulated as a set prediction problem and evaluated on the publicly available dataset [1] in both good and good-bad weather conditions. The proposed method's efficacy is extensively evaluated using the COCO evaluation metric, and the best-proposed model surpasses its state-of-the-art counterpart method by 12.55% and 12.48% in both good and good-bad weather conditions.**

I. INTRODUCTION

The research and technological advancements made in the past three decades have enabled autonomous vehicles to become a certainty [2]. The development of sensors and computing technology has resulted in small size and cost-effective hardware for the autonomous vehicle. The sensor's accuracy, resolution, and latency have improved over the years to be consolidated in autonomous vehicles. The dynamic perception of the environment is fundamental to the safety of the autonomous vehicle. A broad set of exteroceptive sensors is utilized in the autonomous vehicle's hardware suite to perceive the environment. The typical sensor modalities include the camera, Lidar, radar, and sonar help the autonomous vehicle attempt to map, understand, and navigate the environment [3]. Among other perception tasks, the autonomous vehicle's most fundamental task is to identify and understand the objects encompassing it.

The most common exteroceptive sensor modality that is widely deployed for the perception of autonomous vehicles is cameras. Cameras operate in the visible spectrum and provide high-resolution information about the environment, which are essential to identify traffic lights, lane marking, and road obstacles. Besides providing in-depth details of the surroundings, cameras are sensitive to adverse weather conditions, illumination, and sun-glare, resulting in low-level image data that impediment the perception's accuracy for the autonomous vehicles. On the other hand, Lidar provides the 3D information about the environment as a surrogate to 2D details from the cameras. Lidar uses invisible laser light to measure the accurate distance of the objects. It measures nearly thousands of points to develop a 3D point cloud map of the senor's surroundings. Despite providing the 3D information, Lidar has a detriment of being expensive and also sensitive to adverse weather conditions. Contrary to cameras and Lidar, radar is less prone to adverse weather conditions and is used as a sensor modality for detecting small hazard objects in the autonomous vehicle sensor suite. Most of the radar operates on the principle of Doppler's effect by firing radio waves at the target area and analyzing the reflected wave's frequency to estimate the object's speed and position. These type of radars mostly find their application in object tracking for the autonomous vehicle. Besides, there are scanning radars that provide the high resolution $360 \deg$ range-azimuth images and are used for object detection in the perception stack of autonomous vehicles. Moreover, they are relatively cost-effective as compared to Lidar [4].

Object detection forms the basis for the perception module and is formally done by using cameras and Lidar. A lot of research is carried out on the camera-based object detection that includes many state-of-the-art object detectors, for instance, YOLO-v3 [5], SSD [6], Faster-RCNN [7], RefineDet [8], M2Det [9]. These object detectors have limitation to operate in the adverse weather conditions [10]. In addition, much research is also focused on the use of Lidar for object detection. For this purpose, the Lidar data is represented either in rasterized [11] or geometric format [12] [13] but have limitation to processing computation and adverse weather condition. Besides, camera and Lidar object detection, there

---

[1]https://www.daimler.com/innovation/case/autonomous/safety-first-for-automated-driving-2.htm

is still a room of improvement for the radar-based object detection. In this work, radar-based object detection in different weather condition is explored without utilizing any auxiliary sensor modality information. The radar-based object detection is formulated as a set prediction problem by designing a proposed network that includes channel boosting feature ensemble followed by transformer encoder-decoder architecture. The radar's data is complex; to determine rich feature representation, the channel boosting scheme is adopted in which the input radar image is transformed into different color space. The input radar image and color space images are followed by the respective backbone network for the features generation. The respective backbone networks' features are concatenated, and these image features are fed to transformer encoder-decoder architecture for object detection. The proposed network is extensively evaluated using the COCO evaluation metrics. The experimental evaluation shows the efficacy of proposed work with the state-of-the-art method [1]. The main contributions are listed as follows:

1) A novel channel boosting feature ensemble method is designed for the feature representation of radar data.
2) The channel boosting feature ensemble method is used in conjunction with the transformer encoder-decoder architecture for the radar-based object detection.
3) The efficacy of the proposed method (best-model) illustrates 12.55% and 12.48% increase in mean average precision (mAP) score in both good and good-bad weather conditions for radar-based object detection as compared to state-of-the-art (best-model) [1].

The rest of the paper is organized as follows: Section II gives related work. Section III explains the proposed methodology. The experimentation and results are discussed in Section IV, and section V concludes the paper.

II. RELATED WORK

The researchers in [14]–[16] have explored radar-based object detection in the domain of autonomous vehicle. [17] has proposed a post-processing algorithm for object detected by highly accurate and reliable radar. The algorithm clusters multiple detections form a single object and track them using Kalman filter. [18] handles the uncertainties of laser radar by occupancy grid framework. The occupancy grid map is formed by using the inverse sensor model in the cartesian coordinate framework. The local grid map and global grid map, generated by temporal integration of sensor data are fused using the Dempster rule of combination. The conflict information is used to detect the moving object.

The fusion of radar and camera data for object detection is done in [19]–[21]. [22] has proposed the architecture for frontal object detection using radar data and a single camera. The detections are fused and then tracked using the Kalman filter. [23] represents a framework to identify an occupying area by fusing the data from millimetre radar and video camera. [24] performs the radar vision fusion in three steps. First radar guides a selection of the small number of candidates images for processing. Second, sparse representation is

generated by normalizing the selected attention window and is processed by orientation-selective feature detector. Finally, a learned multi-layered network is used to classify the sparse representation of different objects. [25] has proposed a multi-sensor fusion system for object detection and tracking for autonomous vehicles. They have utilized the camera, lidar and radar to perceive the environment and develop a system to detect pedestrian, bicyclist and moving objects. The recognition framework helps to improve tracking and data association of the detected objects by multi-sensor data.

The aforementioned algorithm uses traditional techniques for object detection. Since the advent of the convolutional neural network (CNN), object detection accuracy has dramatically increased. CNN learns informative feature representation of objects contributes to success. In literature, convolutional neural networks are used for radar-based object detection. [26] has proposed a two-step object detection framework called radar region proposal network. The region proposal network generates a proposal by mapping radar data onto images and generates a set number of anchor boxes used to detect objects and calculate their distance from the vehicle. [27] introduces an architecture for object detection that utilizes RGB camera images and radar sensor data. They have used radar data to identify high confidence point in the image to develop better feature representation and generate proposals for object detection. [1] has proposed radar dataset in adverse weather conditions and detect vehicles using radar data. They have used Faster-RCNN architecture for object detection with two modification. First, they fixed the anchor size and second, the bounding box has an additional parameter angle.

A new paradigm of the object detection algorithm has recently evolved based on a new attention-based building block called transformer network. The attention mechanism in transformers aggregates information of the whole input sequence, which give them the advantage of global computations and perfect memory. [28] introduces an encoder-decoder transformer framework for object detection based on a direct set prediction which bypasses anchor generation in typical object detection network.

III. PROPOSED METHOD

Feature extraction plays an imminent role for many deep-learning-based object detection task. In this work, a channel boosting feature ensemble method along with a direct set prediction for object detection using an encoder-decoder transformer network is proposed for the radar-based object detection. Fig.1 illustrates the architecture of the proposed work.

A. Channel Boosting Feature Ensemble Method

The radar data exhibits large variations and thus requires a robust feature extraction mechanism. In this study, a channel boosting in conjunction with the ensemble method is used to extract features for radar images. First, the channel boosting method is used by aggregating the input image's RGB channels by converting it to LUV and CIELAB color
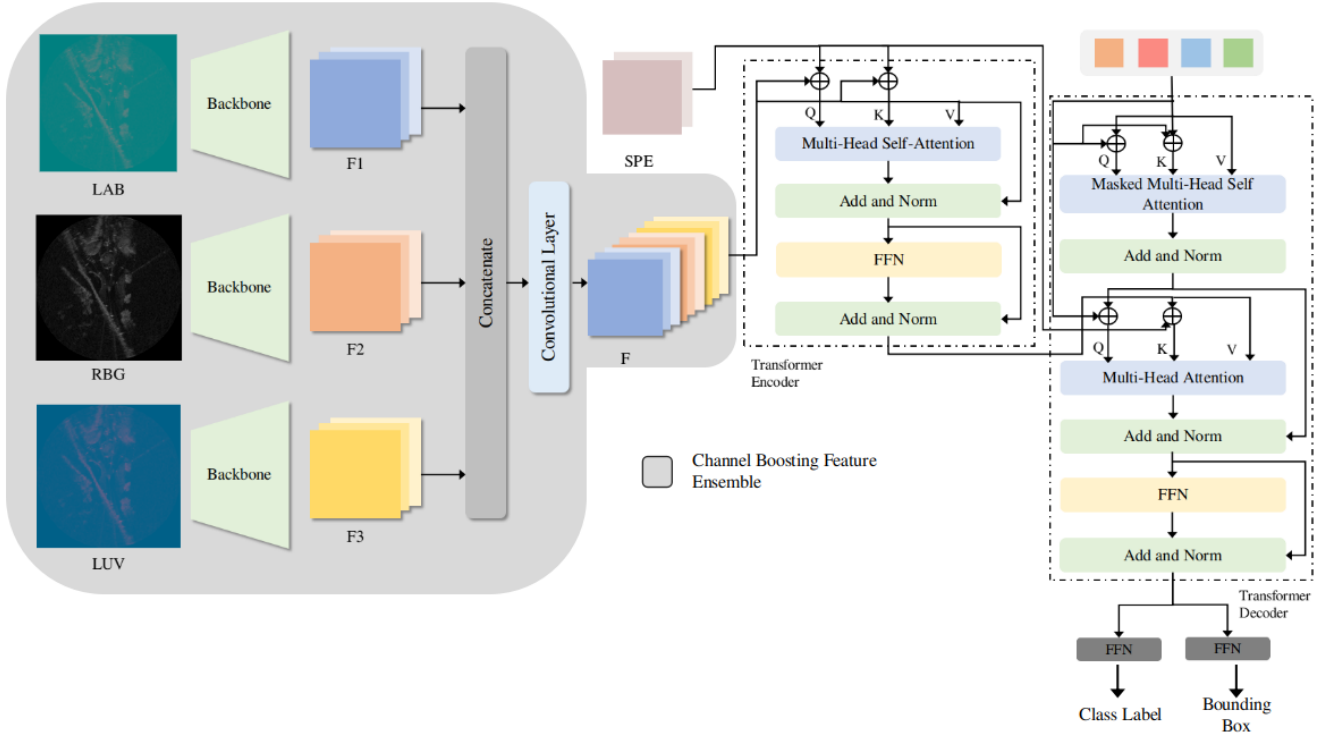
Fig. 1: The overall framework for the proposed method. It includes a channel boosting feature ensemble method along with transformer encoder-decoder network for the radar-based object detection.

spaces. The conversion to these color spaces enables the channel boosting by providing a good representation of the input image and helps the network to learn the modular and hierarchical representation of the image. Fig.1 illustrates the channel boosting ensemble method for the representation of the input image. Suppose $x_i \in \mathbb{R}^{D \times H \times W}$ for $i \in I = (RGB, LUV, CIELAB)$ is the input image with $x_{i_{rgb}}$, $x_{i_{luv}}$ and $x_{i_{lab}}$ corresponds to the variants of the input image. Each channel variant of input image (RGB, LUV, and CIELAB) is followed by the convolution neural network (CNN) backbone for feature representation. Suppose the $x_{i \in rgb}$ with $D = 3$ (initially), is passed to the backbone for the generation of low-resolution activation map $f = \mathbb{R}^{C \times H \times W}$ with typical values of $C = 2048$. Similarly, the other two color space variants of the input image are processed by the backbone network resulting in low-resolution feature maps of each dimension of $C = 2048$ respectively. These low-resolution feature activation maps are concatenated and then followed by an additional convolutional layer to produce the feature activation maps from the channel boosting ensemble method with a dimension of $C = 2048$. The reason for this additional convolutional layer is empirical and also to reduce the computation cost.

B. Transformer Encoder-Decoder

1) Encoder: The transformer encoder expects a sequence as inputs, so the set of activation maps f produced by channel boosting feature ensemble is reduced to lower dimension space from C to k by applying a 1×1 convolution layer and flattened

to create the new feature map $\tilde{f} \in \mathbb{R}^{k \times HW}$ and is given to the transformer encoder. For each encoder layer, a multi-head self-attention module and the feed-forward network is incorporated in the architecture.

For the multi-head self-attention module's brevity, the explanation of the single head attention mechanism is essential [28] [29]. The query and key-value pair form the basis of the attention and are utilized to map it to the output. The query, key and value are all represented in the vector format. The weight tensor $W_t \in \mathbb{R}^{3 \times \hat{k} \times k}$ for a single attention head $A(X_q, Xk_v, W_t)$ computes the query, key and value embedding after adding the query and key positional encoding ($P_q \in \mathbb{R}^{k \times N_q}$ and $P_{kv} \in \mathbb{R}^{k \times N_{kv}}$) respectively, as shown in Eq.1

$$[Q:K:V] = [W_{t1}(X_q + P_q) : W_{t_2}(X_{kv} + P_{kv}) : W_{t_3}X_{kv}] \tag{1}$$

where, [:] represents the concatenation, and (W_{t_1}, W_{t_2} and W_{t_3}) are concatenation of W_t. X_q corresponds to the query sequence of length N_q. The key-value sequence is denoted by X_{k_v} of length N_{k_v}. Eq.2 illustrates the computation of activation weights δ by computing the soft-max function of the dot product of queries and keys.

$$\delta_{(i,j)} = \frac{\exp^{\frac{1}{\sqrt{k}} Q_i^T K_j}}{\sum_{j=1}^{N_{kv}} \exp^{\frac{1}{\sqrt{k}} Q_i^T K_j}} \tag{2}$$

The positional embedding is learned and shared across all the attention layers for the given query, key-value pair [28]. The

final output of the single head attention module is computed by Eq.3.

$$A(X_q, X_{kv}, W) = \sum_{j=1}^{N_{kv}} \delta_{ij} V_j \qquad (3)$$

The multi-head self-attention module is the concatenation of M single head attention modules followed by the projection matrix L. The projection matrix is the combination of residual connection, dropout and layer normalization. The overall computation is described in Eq.4. In the case of multi-head attention module the only difference in computing the attention weight as described in Eq.2 is the change of scale fraction k to $\hat{k} = \frac{k}{M}$.

$$\begin{aligned} X_q{}' &= [A(X_q, X_{kv}, W_1) : ... : A(X_q, X_{kv}, W_M)] \\ \tilde{X}_q &= \psi(X_q + dropout(LX_q{}')) \end{aligned} \qquad (4)$$

2) Transformer Decoder: The decoder follows the same structure of sub-layer as the encoder. It constitutes of two multi-head self-attention modules used for transforming the N embedding size of k.

The decoder's input are queries that are initially set to zero, N object queries that are learnt positional encodings, and the encoder memory. In the decoder, the positional encodings are added to each attention layers. The decoder's aforementioned inputs are used to produce the final set of prediction that includes bounding box and class labels through multi-head self-attention decoder modules. It is to be mentioned here that the first self-attention layer in the first decoder layer is skipped in the processing of decoding the embeddings.

C. Feed-Forward Network (FFN)

The decoder's output embedding is fed to the feed-forward neural network for the prediction of class labels and bounding box. A 3 layer perceptron network with ReLU activation, a hidden dimension of k and a linear projection layer is used for the final prediction output. The output of FFN consists of height, the width of the bounding box w.r.t image, normalized center coordinates and class labels. Besides, an extra no-class label is also used for no object detected in the input image. In this work, the number of classes is 2 (vehicle and no-vehicle) because the intra-class object detection will not provide any sufficient information because of the radar data nature.

IV. EXPERIMENTATION AND RESULTS

A. Dataset

RADIATE (RAdar Dataset In Adverse weaThEr) dataset is used in this study [1]. The motivation to use RADIATE is to facilitate the object detection research in adverse weather conditions and understand the dynamic environment better so that an autonomous vehicle can safely navigate. The dataset uses Navtech CTS350-X radar to collect the data. The radar renders $360°$ high-resolution range-azimuth images. It has a maximum operating range of 100 meters and 0.175m range resolution, $1.8°$ azimuths and elevation resolution. The dataset is collected in different weather conditions including sunny,

TABLE I: The RADIATE dataset partition topology for training, and testing the proposed method in adverse weather conditions

	No. of Images	No. of vehicles
Train Set (Good Weather)	23091	106931
Train Set (Good and Bad Weather)	9760	39647
Test Set	11289	147005

overcast, night, snow, fog and rain. The dataset consists of 3 hours of annotated radar images with an estimated 200K labels vehicles. The driving scenarios covers urban, motorway and suburban driving. The images are labelled using 2D bounding boxes. Each bounding box define $(x, y, width, height, angle)$, where (x, y) are the top-left pixel location and angle define counter-clockwise rotation. The dataset is divided into three parts, training in good weather, training in good-bad weather and test set that contain both good and bad weather. We have adopted standard partition as proposed by [1] for a fair comparison, shown in Table I.

B. Evaluation Metric

Numerous evaluation metrics exist to evaluate the accuracy of object detection in the images. Here, we have used a standard MS COCO evaluation metric [30]. IoU (Intersection over Union) defined the intersection between the area of ground-truth bounding box and area of the predicted bounding box shown by Eq. 5. The prediction is True Positive (TP) if $IoU > threshold$ and False Positive (FP) if $IoU < threshold$.

$$IoU = \frac{A_p \cap A_{gt}}{A_p \cup A_{gt}} \qquad (5)$$

Recall and Precision are calculated using Eq. 6.

$$Recall = \frac{TP}{TP + FN}, Precision = \frac{TP}{TP + FP} \qquad (6)$$

The Average Precision(AP) is calculated per class. AP is the area under the PR curve, shown by Eq. 7.

$$AP[class] = \frac{1}{\#thresh} \sum_{IoU \in thresh} AP[class, IoU] \qquad (7)$$

COCO evaluation uses a threshold range from 0.5 to 0.95 with a step size of 0.05. In the experimentation, $mAP_{IoU=0.5}$ is utilized, which is also similar to PASCAL VOC [31] evaluation.

C. Experimentation

The proposed network is trained on two GPUs having $24Gb$ memory in total using Pytorch deep learning library. The input training data is scaled, augmented, randomly horizontal flipped, randomly resized and randomly cropped in order to provide the network to learn the global relationship. The training process is run for 125 epochs having the learning rate drop of factor 10 at 100 epochs. The Adam (Adaptive Momentum Estimation) is used as an optimizer with 10^{-5} and 10^{-4} learning rates for the backbone and transformer network,
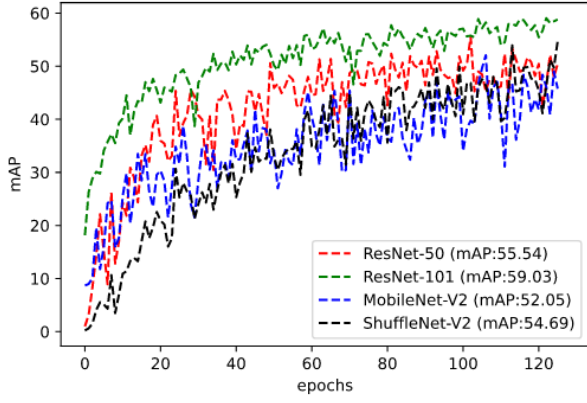
Fig. 2: The illustration of mAP scores for all the models used as a backbone network in the proposed method trained on good-bad weather conditions and tested on test data. ResNet-101 has the best results.
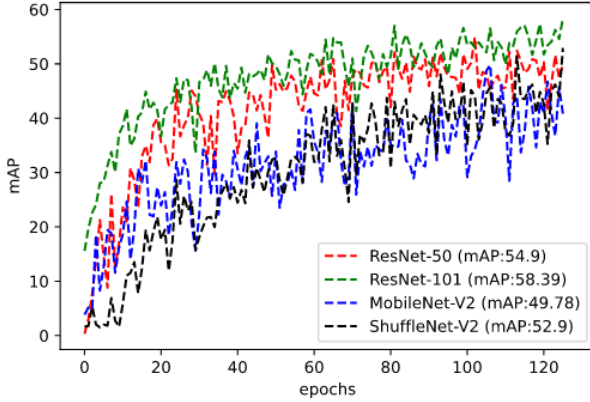

Fig. 3: The illustration of mAP scores for all the models used as a backbone network in the proposed method trained on good weather conditions and tested on test data. ResNet-101 has the best results.

respectively. The pre-trained backbone network is imported from the Torchvision with frozen the batch-norm layer and discarding the last classification layer. In the experimentation, the ResNet-50 [32], RestNet-101 [32], MobileNet-V2 [33] and ShuffleNet-V2 [34] are used as a backbone networks. In the transformer network, the number of encoders and decoders are equal to 6. All the transformer weights are randomly initialized with Xavier initialization. A dropout of 0.1 is used after every multi-head attention and feed-forward network layer before normalization. For the spatial positional encodings, an absolute position encodings are used that are function of sine and cosine with different frequencies and are concatenated to get the final position encoding across the k channel.

Since, the proposed method infers the set of N predictions, in this work, inspired by [28] an optimal bipartite matching scheme is adopted for the loss calculations. Eq.8 illustrates the bipartite matching loss function between $\hat{y}$ set of predictions N added with no object class and ground-truth (y) also padded

with the set of size N.

$$\tilde{\eta} = \underset{\eta \in C_N}{argmin} \sum_{j}^{N} L_{match}(y_j, \hat{y}_{\eta(j)}) \qquad (8)$$

where C_N corresponds to the permutations of N elements. $L_{match}(y_j, \hat{y}_{\eta(j)})$ illustrates the pair-wise matching cost between ground-truth and prediction. A Hungarian algorithm is used to compute this assignment problem between ground-truth and prediction. The matching cost constitutes towards both the class labels and the bounding box between the ground-truth and the predictions. Suppose each element j of ground-truth set is represented as $y_j = (c_j, b_j)$ where c_j is the class label and $b_j \in [0,1]^5$ is the bounding box vector corresponds to normalized center coordinates, height, width and angle for the respective image. For the prediction, the class probability and bounding box is defined as $\hat{p}_{\eta j}(c_j)$ and $\hat{b}_{\eta j}$ respectively. The matching between ground-truth and prediction is done by finding the direct one-to-one correspondence without duplicates. The modified Hungarian loss function with angle inclusion in the bounding box is illustrated in Eq.9 using the aforementioned notations of the class label and bounding box.

$$L_{Hungarian}(y, \hat{y}) = \sum_{j=1}^{N} [-log\hat{p}_{\eta(j)}(c_j) + \qquad (9)$$
$$\mathbf{1}_{c_j \neq (\phi)} L_{box}(b_j, \hat{b}_{\hat{\eta}}(j))]$$

where $\hat{\eta}$ corresponds to optimal assignment. The $L_{box}(.)$ in the Hungarian loss function is used for scoring the bounding box, and in comparison to other detectors the box prediction is computed directly. In other to compensate the scaling issue, the l_1 loss with the complete IoU loss is used. The complete IoU [35] loss is illustrated by the Eq.10.

$$L_{CIoU} = 1 - IoU + \frac{\rho(\mathbf{b}, \boldsymbol{b_{gt}})}{c^2} + \alpha \nu \qquad (10)$$

where $1 - IoU + \frac{\rho(\mathbf{b}, \boldsymbol{b_{gt}})}{c^2}$ corresponds to distance IoU with $\mathbf{b}$, $\boldsymbol{b_{gt}}$ denotes the bounding box points of predicted and ground-truth respectively. ρ denotes the euclidean distance and c illustrates the diagonal length of smallest enclosing box that covers the two boxes. α is the positive trade-off parameter and the ν is used to compute the consistency of aspect ratio defined by Eq.11 and Eq.12 respectively.

$$\alpha = \frac{\nu}{(1 - IoU) + \nu} \qquad (11)$$

$$\nu = \frac{4}{\pi^2}(arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w}{h})^2 \qquad (12)$$

The $L_{box}(.)$ is represented as shown in Eq.13

$$L_{box}(b_j, \hat{b}_{\hat{\eta}}(j)) = \lambda_{iou} L_{CIoU}(b_j, \hat{b}_{\hat{\eta}}(j)) + \qquad (13)$$
$$\lambda_{l1} \left\| b_j - \hat{b}_{\hat{\eta}}(j) \right\|$$

where $\lambda_{l1} = 4$ and $\lambda_{iou} = 2$ are used in training the proposed method.
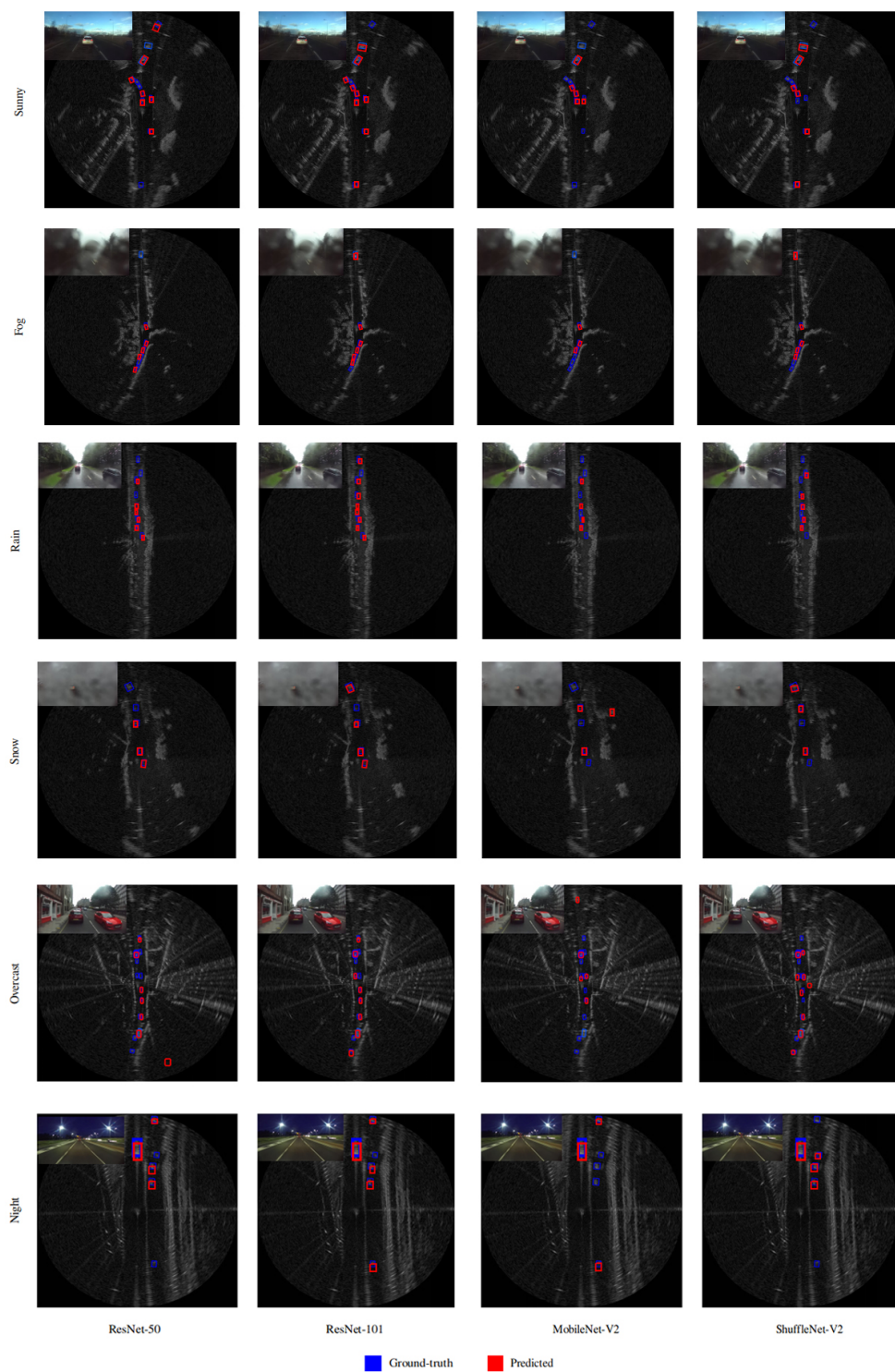
Fig. 4: The qualitative results of proposed radar-based object detection in different weather conditions with all the four backbone networks used along with transformer encoder-decoder network is illustrated.

TABLE II: mAP Score on Radar Data with Good and Bad Weather Conditions

Models	mAP Score
ResNet-50 Trained [1]	45.77
ResNet-101 [1]	46.55
Ours (Backbone: ResNet-50) (Baseline)	55.54
Ours (Backbone: ShuffleNet-V2)	54.69
Ours (Backbone: MobileNet-V2)	52.05
Ours (Backbone: ResNet-101)	**59.03**

TABLE III: mAP Score on Radar Data with Good Weather Conditions

Models	mAP Score
ResNet-50 Trained [1]	45.31
ResNet-101 [1]	45.84
Ours (Backbone: ResNet-50) (Baseline)	54.90
Ours (Backbone: ShuffleNet-V2)	52.90
Ours (Backbone: MobileNet-V2)	49.78
Ours (Backbone: ResNet-101)	**58.39**

For the quantitative and qualitative evaluation, a baseline network of proposed work is trained on two sets of datasets that include good-bad weather conditions and only good weather conditions. The proposed baseline network is compared with [1] for the evaluation to measure the mean average precision (mAP) scores. Table-II and Table-III illustrates the mAP score of proposed baseline method on testing data. In addition, to further explore the proposed method's effectiveness, the proposed method is trained using RestNet-101, MobileNet-V2 and ShuffleNet-V2 as a backbone network. In the quantitative analysis, ResNet-101 has outperformed the other backbone network mAP scores and, also surpasses the [1] scores by a margin of 12.48% in good-bad weather conditions and by 12.55% in good weather conditions with its counterpart network. Fig.2 and Fig.3 illustrates the running mAP scores of good-bad weather conditions and good weather conditions across the number of epochs on the testing data. Fig.4 shows the qualitative results of radar object detection of all the backbone used in the experimentation in different weather conditions along with the accompanying optical image of the scene.

V. Conclusion

This study focuses on radar-based object detection in adverse weather condition for autonomous vehicles. A novel architecture has introduced which utilizes the channel boosting feature ensemble for more desirable feature extraction. The extracted features are then concatenated and feed to the encoder-decoder transformer network, which uses a direct set prediction method for object detection. The efficacy of the proposed algorithm is evaluated using standard COCO evaluation. The mean average precision of 59.03% and 58.39% are achieved on test data for the proposed method with ResNet-101 backbone netowrk trained on good-bad weather conditions and good weather conditions respectively.

In future work, we aim to fuse Lidar data with radar to improve object detection in adverse weather condition. Moreover, use image data to classify the weather condition to optimize the perception system of the autonomous vehicle.

References

[1] M. Sheeny, E. De Pellegrin, S. Mukherjee, A. Ahrabian, S. Wang, and A. Wallace, "Radiate: A radar dataset for automotive perception," *arXiv preprint arXiv:2010.09076*, 2020.

[2] K. Bimbraw, "Autonomous cars: Past, present and future a review of the developments in the last century, the present scenario and the expected future of autonomous vehicle technology," in *2015 12th international conference on informatics in control, automation and robotics (ICINCO)*, vol. 1. IEEE, 2015, pp. 191–198.

[3] S. Campbell, N. O'Mahony, L. Krpalcova, D. Riordan, J. Walsh, A. Murphy, and C. Ryan, "Sensor technology in autonomous vehicles: A review," in *2018 29th Irish Signals and Systems Conference (ISSC)*. IEEE, 2018, pp. 1–4.

[4] F. Rosique, P. J. Navarro, C. Fernández, and A. Padilla, "A systematic review of perception system and simulators for autonomous vehicles research," *Sensors*, vol. 19, no. 3, p. 648, 2019.

[5] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[7] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks. corr abs/1506.01497 (2015)," *arXiv preprint arXiv:1506.01497*, 2015.

[8] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4203–4212.

[9] Q. Zhao, T. Sheng, Y. Wang, Z. Tang, Y. Chen, L. Cai, and H. Ling, "M2det: A single-shot object detector based on multi-level feature pyramid network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 9259–9266.

[10] S. D. Pendleton, H. Andersen, X. Du, X. Shen, M. Meghjani, Y. H. Eng, D. Rus, and M. H. Ang, "Perception, planning, control, and coordination for autonomous vehicles," *Machines*, vol. 5, no. 1, p. 6, 2017.

[11] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4490–4499.

[12] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[13] S. Azam, F. Munir, A. Rafique, Y. Ko, A. M. Sheri, and M. Jeon, "Object modeling from 3d point cloud data for self-driving vehicles," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 409–414.

[14] C. C. Schwesig and I. Poupyrev, "Radar-based object detection for vehicles," Oct. 29 2019, uS Patent 10,459,080.

[15] D. M. Gavrila, "Sensor-based pedestrian protection," *IEEE Intelligent Systems*, vol. 16, no. 6, pp. 77–81, 2001.

[16] S. Miyahara, "New algorithm for multiple object detection in fm-cw radar," SAE Technical Paper, Tech. Rep., 2004.

[17] A. Manjunath, Y. Liu, B. Henriques, and A. Engstle, "Radar based object detection and tracking for autonomous driving," in *2018 IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM)*. IEEE, 2018, pp. 1–4.

[18] J. Duan, L. Ren, L. Li, and D. Liu, "Moving objects detection in evidential occupancy grids using laser radar," in *2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, vol. 2. IEEE, 2016, pp. 73–76.

[19] M. Gong, Z. Zhou, and J. Ma, "Change detection in synthetic aperture radar images based on image fusion and fuzzy clustering," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 2141–2151, 2011.

[20] Y.-T. Zhou, "Multi-sensor image fusion," in *Proceedings of 1st International Conference on Image Processing*, vol. 1. IEEE, 1994, pp. 193–197.

[21] K. Abe, S. Tokoro, and K. Suzuki, "Object detection system and method of detecting object," Apr. 15 2008, uS Patent 7,358,889.

[22] R. O. Chavez-Garcia, J. Burlet, T.-D. Vu, and O. Aycard, "Frontal object perception using radar and mono-vision," in *2012 IEEE Intelligent Vehicles Symposium*. IEEE, 2012, pp. 159–164.

[23] T. Kato, Y. Ninomiya, and I. Masaki, "An obstacle detection method by fusion of radar and motion stereo," *IEEE transactions on intelligent transportation systems*, vol. 3, no. 3, pp. 182–188, 2002.

[24] Z. Ji and D. Prokhorov, "Radar-vision fusion for object classification," in *2008 11th International Conference on Information Fusion*. IEEE, 2008, pp. 1–7.

[25] H. Cho, Y.-W. Seo, B. V. Kumar, and R. R. Rajkumar, "A multi-sensor fusion system for moving object detection and tracking in urban driving environments," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 1836–1843.

[26] R. Nabati and H. Qi, "Rrpn: Radar region proposal network for object detection in autonomous vehicles," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 3093–3097.

[27] R. Yadav, A. Vierling, and K. Berns, "Radar+ rgb attentive fusion for robust object detection in autonomous vehicles," *arXiv preprint arXiv:2008.13642*, 2020.

[28] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *arXiv preprint arXiv:2005.12872*, 2020.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[30] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[31] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition. corr abs/1512.03385 (2015)," 2015.

[33] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[34] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.

[35] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-iou loss: Faster and better learning for bounding box regression." in *AAAI*, 2020, pp. 12 993–13 000.