NeRF-FCM: Attention-based Feature Calibration Mechanisms for 3D Object Detection Using NeRF

Hana Lebeta Goshu*, Jun Xiao*, Kin-Chung Chan*, Cong Zhang*, Mulugeta Tegegn Gemeda† and Kin-Man Lam*
* The Hong Kong Polytechnic University, Hong Kong
{hana-lebeta.goshu, jun.xiao, alfred.chen, cong-clarence.zhang}@connect.polyu.hk, enkmlam@polyu.edu.hk
† Jimma Institute of Technology, Ethiopia
mulugeta.geneda@ju.edu.et

***Abstract*—With the fast development of 3D vision, 3D object detection based on posed RGB images has become increasingly popular and attracted significant attention from researchers in recent years. Given the remarkable performance of Neural Radiance Field (NeRF) in modeling 3D scenes, recent 3D detection methods utilizing posed RGB images generated by NeRF models have achieved promising results. However, NeRF-based models often suffer from poor generalization and are prone to generating inconsistent image content for unseen views, which inevitably degrades the performance of existing NeRF-based 3D detectors. In this paper, we propose an effective feature calibration method to enhance the performance of 3D detection models based on posed RGB images produced by NeRF models. Specifically, our proposed method efficiently recalibrates the 3D features extracted from the backbone network, and adaptively computes the weights for fusion based on the statistical properties of the features. Experiments show that our method significantly outperforms the baseline model, achieving improvement of +8.6 AP@0.5, +5.5 AP@0.5, and +5.1 AP@0.5 on the Hypersim, 3D-FRONT, and ScanNet benchmarks, respectively, with anchor-free heads. Particularly, compared with the baseline model, our method can more accurately predict the 3D bounding boxes in 3D space, even when objects are poorly reconstructed by NeRF while keeping low computational costs with a minimal increase in model complexity.**

***Index Terms*—3D Object Detection, NeRF, Channel Attention, Multi-View**

I. Introduction

3D object detection is a foundational topic in the 3D vision tasks, attracting significant attention from researchers in recent years due to its substantial industrial value in applications such as robotics [1], autonomous driving [2], Virtual Reality (VR) [3], Augmented Reality (AR) [4], etc. Most existing 3D object detection methods rely on point clouds [5]–[8] and depth images [9] because these data types provide precise geometric information. However, obtaining highly accurate geometric data typically requires costly 3D sensors, including laser scanners, depth cameras, and stereo cameras, which can be impractical for real-world applications. To reduce high costs, 3D object detection methods based on posed RGB images have been proposed [10], [11]. These methods leverage monocular techniques and have shown promising results. However, monocular-based approaches are inevitably degraded when faced with limited fields of view, occlusion, and scale uncertainty. To address these issues, ImVoxelNet [12] adopts a multi-view image for 3D object detection, effectively learning feature representations of 3D voxel volumes from multi-view 2D images. Nonetheless, this method struggles with severe occlusion and complex geometric information.

Recently, Neural Radiance Field (NeRF) [13] has demonstrated remarkable capabilities in 3D scene representation and novel view synthesis. Inspired by this, several 3D object detection methods based on NeRF have been proposed, which achieve promising performance [14]–[17]. For example, NeRF-RPN [14] predicts the location of objects in 3D scenes using posed RGB images rendered by NeRF. Specifically, it takes NeRF outputs (i.e., color and density information) as input to produce 3D bounding boxes. During feature extraction, the feature maps received by the Region Proposal Network (RPN) head are highly biased towards capturing local contexts, neglecting global information. This leads to regions of interest that rely solely on local spatial information, further deteriorating detection accuracy in complex 3D scenes.

In this paper, we propose a robust method to incorporate global information into NeRF-based 3D detection methods, thereby enhancing detection performance. In the human visual system, global context (e.g., surroundings and background) is crucial for object recognition, but previous methods have predominantly focused on local features. In contrast, our method effectively fuses global and local features by employing a 3D Global Average Pooling (GAP) mechanism for a channel-wise attention mechanism. Additionally, inspired by ECA-Net [18], we propose an efficient method to learn the inter-dependency between global and local features with negligible additional cost. Consequently, our method can adaptively aggregate global and local features, learning effective 3D feature representations to enhance the accuracy of 3D object detection.

The contributions of this paper are summarized as follows:

- We demonstrate that incorporating global information can enhance the quality of NeRF-based 3D object detection from posed RGB images.
- We propose a novel framework for NeRF-based 3D object detection that utilizes a channel attention module to improve 3D feature representations and produce accurate 3D bounding boxes with minimal computational costs.
- Our experiments show that our proposed method can significantly enhance 3D object detection accuracy and suppress an existing baseline by +8.6 AP@0.5, +5.5 AP@0.5,

and +5.1 AP@0.5 on the Hypersim, 3D-FRONT, and ScanNet indoor datasets, respectively.

II. Related Works

A. Point Cloud-based 3D Object Detection

3D object detection from point clouds has achieved great results in outdoor and indoor scenes because point clouds provide reliable geometric structural information. It is divided into two types: grid-based and point cloud-based approaches. Grid-based approaches, such as VoxelNet [5] and PointPillars [6], typically transform point clouds into 3D grids and use 3D CNNs to process the grid representation. However, this approach is memory-intensive due to the dense volumetric representation and the use of 3D CNNs. To mitigate this problem, sparse fully-convolutional detection methods, such as FCAF3D [19] and [20], have been introduced, improving the quality of 3D object detection in both accuracy and efficiency. Conversely, point cloud-based methods detect 3D objects from the point clouds [7], [8]. These methods utilize deep Hough voting, as proposed in VoxelNet [7], to predict bounding box parameters from point features. However, voting-based 3D detection methods pose challenges when used with reconstruction networks, since they need extra data annotation on the point clouds. Recent point cloud-based methods have attracted significant attention owing to their strong performance in 3D detection, particularly with the emergence of various 3DGS methods [21], [22], eliminating the need for expensive 3D sensors. Methods that incorporate multi-views to enhance scene understanding have also become more prominent. ImVoxelNet [12] has achieved impressive results in indoor 3D detection using the 3D voxel-based feature volume [23]. Nevertheless, it fails to retain the inherent geometry of input scenes while constructing the feature volume.

B. Channel Attention and Object Detection

In the field of computer vision, channel attention methods are essential for improving the performance of deep CNNs across various tasks [18], [24]–[27]. The main challenge for convolutional networks is their limited ability to extract global features, as the convolution kernels mainly focus on local regions in space. To address this issue, channel attention methods have been investigated and widely used to enhance the representational power of networks by aggregating global information. The Squeeze-and-Excitation (SE) block in a SE-Net [24] has become a paradigm for channel attention, which assigns learned weights to different channels of convolutional layers and recalibrates the feature maps by capturing channel-wise dependencies, thus improving detection performance. CBAM [25] uses convolutions with large-size kernels to encode spatial information. ECA-Net [18], based on SE-Net, employs 1D convolutions to capture local cross-channel interaction while resolving the negative effects of channel dimension reduction. However, these techniques are mainly used for 2D object detection. To improve the quality of 3D object detection using NeRF from posed RGB images, this paper presents the 3D Squeeze and Excitation Channel Attention (3D SE-CA) and 3D Efficient Channel Attention (3D ECA) modules.

III. Methods

A. Overview

The architecture of our proposed method is illustrated in Fig. 1. Following NeRF-RPN [14], our method uses a pre-trained NeRF model and uniformly samples the volume density and RGB color information to generate a feature volume. To extract the feature maps, we employ a 3D backbone network. Before passing the extracted feature maps to the Feature Pyramid Network (FPN) [28], we integrate channel attention mechanisms to enhance feature representation. Finally, the RPN head uses the feature pyramid to produce a set of regional proposals.

B. Sampling Input and Extracting Features

To obtain accurate 3D geometric information, we leverage the capabilities of a pre-trained NeRF model. Thus, our network directly produces 3D bounding boxes of Regions of Interest (ROIs) in 3D space by using the extracted 3D volumetric information from NeRF as input. The extracted radiance and density volume field are uniformly sampled in order to build a feature volume. Each sample can be represented as (r, g, b, α), where c = (r, g, b) represents the emitted averaged color and α, the opacity, can be obtained as follows [14]:

$$\alpha = \text{clip}\left(1 - \exp(-\rho \cdot \delta), 0, 1\right), \tag{1}$$

where ρ is the volume density and δ represents the distance between two adjacent points, set to 0.01.

We employ a 3D backbone network to extract feature maps after sampling inputs on the 3D grid and extracting RGB and volume densities. Subsequently, we apply our 3D channel attention methods to further extract channel-wise global information and recalibrate the weight of each channel to boost feature representation. We embed 3D SE-CA and 3D ECA between the 3D backbone and 3D FPN to investigate the effect of adding these channel attention methods and determine the optimal placement for best performance.

The 3D FPN takes enhanced feature maps from the channel attention module to produce multi-scale feature volumes, addressing scale variation issues in images and improving the ability of the model to accurately detect objects of various sizes. The 3D FPN exploits all layers of the backbone network, i.e., $C2, C3, C4$, and $C5$, to produce a feature pyramid comprising $P2, P3, P4$, and $P5$. We apply the channel attention module to the output of each bottom-up feature map, i.e., $C2, C3, C4$, and $C5$, before passing them to each top-down feature map, i.e., $P5, P4, P3$, and $P2$, as illustrated in Fig. 1. Thus, the $P2, P3, P4$, and $P5$ feature maps acquire global information from the channel attention module, which is then passed to each layer of a top-down pathway of the 3D FPN. Finally, the 3D RPN takes the feature pyramid produced by the 3D FPN block as input to generate 3D bounding boxes.

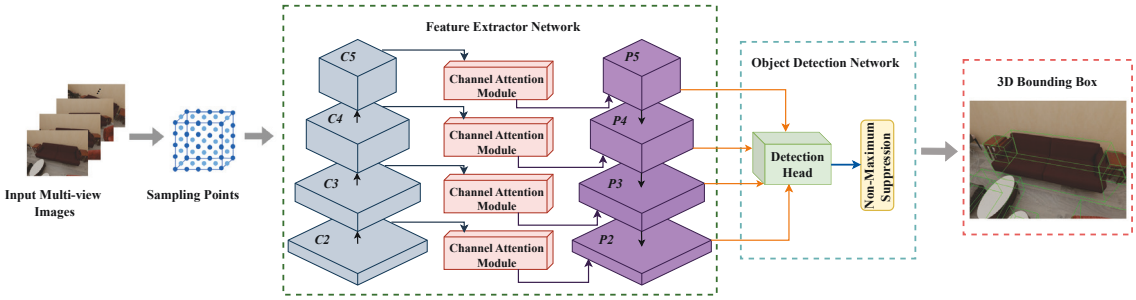

Fig. 1. Overview of our proposed method. Given multi-view images, NeRF retrieves 3D geometric information to improve the 3D detection pipeline. We use VGG19 to extract 3D feature maps and apply a channel attention module on each feature to learn channel importance with global information. The 3D FPN takes both global and local features to produce a feature pyramid, which is subsequently utilized by the detection head to produce accurate 3D bounding boxes.

C. 3D Squeeze and Excitation Channel Attention Module

The proposed 3D SE-CA module extends the 2D SE blocks presented in [24] to a 3D version (see Fig. 2), focusing on channels with more informative features in the 3D features map to improve 3D object detection performance. In the SE block, the transformation F_{se} is divided into the squeeze operation F_{sq} and the excite operation F_{ex}. In F_{sq}, a $1 \times 1 \times 1 \times C$ sized feature map is generated to obtain channel-wise global information statistics while keeping the number of channels C constant using 3D GAP. The output is produced by squeezing input U through the spatial dimensions of $D \times H \times W$ and indicated by $z_i = F_{sq}(u_i)$, which belongs to $Z \in R^C$. The F_{sq} is expressed as follows:

$$z_i = \frac{1}{(D \times H \times W)} \sum_{i=1}^{D} \sum_{j=1}^{H} \sum_{k=1}^{W} u_i(i,j,k), \tag{2}$$

where D, H, and W denote the depth, height, and width, respectively. We set $U = F_{tr}(X)$, where X is an input feature and $U = [u_1, u_2, \ldots, u_c]$, $u_i \in R^{D \times H \times W}$, is the output feature. The channel-wise output of F_{sq} modulates the inter-dependencies of all channels via excitation. The F_{ex} receives Z and learns the inter-channel interactions leveraging two fully connected layers, as follows:

$$s_i = \sigma(W_2 \delta(W_1 Z)), \tag{3}$$

where σ represents sigmoid activation function, δ denotes ReLU function [29], $W_1 \in R^{\frac{C}{r} \times C}$ includes parameters for dimensionality-reduction, and $W_2 \in R^C \times \frac{C}{r}$ includes parameters for dimensionality-increasing layers and $r = 16$. A channel-wised multiplication F_{scale} between each feature map u_i and excitation scalar s_i is ultimately employed to produce final re-scaling feature output U, as follows:

$$U = F_{scale}(u_i, s_i) = u_i \cdot s_i. \tag{4}$$

D. 3D Efficient Channel Attention Module

We extend the original 2D ECA-Net [18] to the 3D version and integrate it with NeRF-RPN [14] for 3D object detection. This module captures local cross-channel interactions while

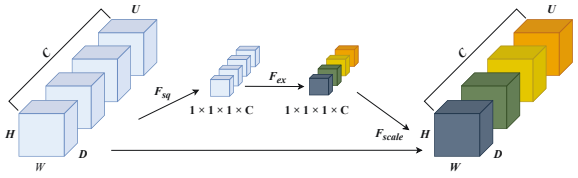

Fig. 2. The 3D Squeeze and Excitation Channel Attention Module.

alleviating the negative impact of channel dimension reduction by using a 1D convolution with adaptive kernel sizes, replacing traditional fully connected layers. This process includes GAP, followed by 1D convolution and sigmoid activation to modulate an input feature map (see Fig. 3). The output from a convolution block is denoted as $x \in R^{W \times H \times D}$, where W, H, and D are width, height, and depth, respectively. The feature map is forwarded to a 3D ECA block to enhance the network representation ability by applying the channel attention mechanism. In the 3D ECA module, the 3D GAP of the channel dimension is computed as follows:

$$g(x) = \frac{1}{(HWD)} \sum_{i=1,j=1,k=1}^{H,W,D} x_{i,j,k}, \tag{5}$$

where $g(x)$ represents features after global average pooling. Moreover, channel weights are calculated as follows:

$$\omega = \sigma(W_p g), \tag{6}$$

where σ denotes a sigmoid function and W_p is a parameter matrix for the channel attention, calculated as follows:

$$\begin{bmatrix} w^{1,1} & \ldots & w^{1,p} & 0 & 0 & \cdots & \cdots & 0 \\ 0 & w^{2,2} & \cdots & w^{2,p+1} & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & w^{C,C-p+1} & \cdots & w^{C,C} \end{bmatrix}, \tag{7}$$

where W_p involves $p \times C$ parameters. Based on (7), the weight of g_i is computed as follows:

$$w_i = \sigma\Big(\sum_{j=1}^{p} w^j g^j\Big), \quad g^j \in \Omega_i^p, \tag{8}$$

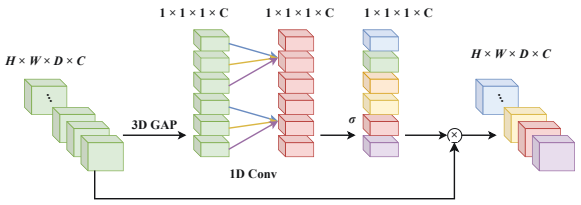

Fig. 3. The 3D Efficient Channel Attention Module.

where Ω_i^p represents the p-set of g_i adjacent channels. Alternatively, the above computation can be performed using $1D$ convolutional operation with the kernel size of p. Thus, the channel features in the g_x are derived as follows:

$$\omega = \sigma\left(\mathrm{C1D}_p(g_x)\right), \tag{9}$$

where C1D denotes one dimension convolution and $p = 3$.

E. 3D Detection Head and Loss Functions

We utilize a 3D RPN head to produce object proposals from the feature pyramid and adopt the loss functions from [14]. We conduct experiments on anchor-free and anchor-based methods to compare the results with the baseline method. We utilize the same loss function as in [30] for the anchor-based method:

$$\begin{aligned} L\left(\{p_i\},\{t_i\}\right) = \frac{1}{N_{cls}} \sum_i L_{cls}\left(p_i, p_i^*\right) \\ + \frac{\lambda}{N_{\text{reg}}} \sum_i p_i^* L_{reg}\left(t_i, t_i^*\right), \end{aligned} \tag{10}$$

where i represents an anchor index, t_i stands for box offsets, and t_i^* denotes the ground-truth bounding boxes. p_i and p_i^* correspond to predicted objectness and ground-truth label, respectively. L_{cls} and L_{reg} denote the classification loss and smooth L_1 loss, respectively, as presented in [31]. N_{cls} and N_{reg} represent the number of anchors utilized in the classification and regression loss computation, respectively, and λ is used to balance these two losses. For anchor-free methods, the centerness loss is added to obtain the final loss function [32]:

$$\begin{aligned} L\left(\{p_i\},\{\boldsymbol{t}_i\},\{c_i\}\right) = \frac{1}{N_{\text{pos}}} L_{cls}\left(p_i, p_i^*\right) \\ + \frac{\lambda}{N_{\text{pos}}} p_i^* L_{reg}\left(\boldsymbol{t}_i, \boldsymbol{t}_i^*\right) + \frac{1}{N_{\text{pos}}} p_i^* L_{ctr}\left(c_i, c_i^*\right), \end{aligned} \tag{11}$$

where N_{pos} represents the number of positive samples, λ denotes balance factor, L_{cls} is the focal loss, L_{reg} is the IoU loss for rotated boxes [33], and L_{ctr} is the centerness loss.

IV. Experiments and Analysis

A. Experimental Setup

In our experiments, we mainly follow NeRF-RPN [14], including backbones, detection head, training recipe, datasets *etc.* Three indoor datasets, Hypersim [34], 3D-FRONT [35], and ScanNet [36], which were constructed as NeRF datasets for 3D object detection by [14], are used. We utilize the NeRF implementation from Instant-NGP [37] to reconstruct them.

Hypersim is the photorealistic synthetic dataset for indoor scene understanding, containing 77,400 images of 461 indoor scenes. For NeRF training in 3D object detection, [14] thoroughly cleaned data based on NeRF reconstruction quality and object annotation usefulness, retaining about 250 scenes.

3D-FRONT is the large-scale dataset with professionally designed synthetic indoor scenes that contain room layouts and textured 3D models of furniture. From the dataset, 159 rooms and their bounding boxes were manually chosen, cleaned, and rendered to generate the NeRF dataset for 3D object detection.

ScanNet is the widely used dataset for indoor 3D detection, containing 2.5 million views from 1513 RGB scans. For each scene, video frames were divided into 100 bins, and bounding boxes were computed from annotated meshes [14]. Models were trained using a given depth and depth-guided NeRF [38].

We use AdamW [39] optimizer with the weight decay of 10^{-3} and the initial learning rate of 3×10^{-4}. For Equations (10) and (11), λ is set to 5.0 and 1.0, respectively. During testing, the top 2,500 proposals are chosen, and redundant proposals are eliminated using Non-Maximum Suppression (NMS), based on rotated IoU with a threshold of 0.1. We train our models using NVIDIA RTX 4090 GPUs with various model configurations on the three datasets. During inference, our method achieves high accuracy with a lower computational cost, with a run time of 61 seconds compared to 67 seconds for the baseline, on the 3D-FRONT dataset. On the Hypersism dataset, the run time of our method is 139 seconds, nearly equivalent to 138 seconds for the baseline on a single GPU, while it significantly enhances 3D object detection performance. We evaluate the methods mainly by average precision (AP) and recall (R) scores with 0.25 IoU and 0.5 IoU thresholds.

B. Experimental Results

We compare our proposed method with existing baseline methods [12], [14], [19], as depicted in Table I. Using VGG19 as the backbone network, with anchor-free and anchor-based RPN heads, we observe that the proposed channel attention methods, 3D SE-CA and 3D ECA, outperform the baseline NeRF-RPN by achieving higher recall scores and AP on all three datasets. This implies that channel attention is beneficial, and our method can effectively enhance 3D object detection performance. Specifically, 3D ECA has made significant improvements compared to the baseline in AP@0.5 on Hypersim (+8.6), 3D-FRONT (+5.5), and ScanNet (+5.1) with anchor-free head. The improvement, particularly in the AP@0.5 evaluation metric, shows that our method is capable of better integrating global information from multi-view and adaptively recalibrating channel weights, thus improving the quality of predicted 3D bounding boxes. Channel-wise interactions have significantly boosted feature representation.

We present visualizations of the predicted 3D bounding boxes generated by NeRF-RPN and our method with anchor-free head, as depicted in Fig. 4. We note that our method provides accurate detection predictions even on objects poorly

TABLE I

QUANTITATIVE COMPARISON WITH EXISTING METHODS ON 3D-FRONT, SCANNET AND HYPERSIM. THE BEST PERFORMANCES OF DIFFERENT EVALUATION METRICS ARE HIGHLIGHTED IN BOLD, WHILE DIFFERENT SETTINGS ARE IN BLUE. R INDICATES A RECALL SCORE AT AN IOU THRESHOLD.

Methods	3D-FRONT				ScanNet				Hypersim			
	R@0.25	R@0.5	AP@0.25	AP@0.5	R@0.25	R@0.5	AP@0.25	AP@0.5	R@0.25	R@0.5	AP@0.25	AP@0.5
ImVoxNet [12]	88.3	71.5	86.1	**66.4**	51.7	20.2	37.3	9.8	19.7	5.7	9.7	2.3
FCAF3D [19]	89.1	56.9	73.1	35.2	90.2	42.4	**67.7**	18.5	47.6	19.4	30.7	8.8
NeRF-RPN (anchor-based)	97.8	76.5	65.9	43.2	88.7	42.4	40.7	14.4	57.1	14.9	11.2	1.3
Ours (3D SE-CA)	**98.5**	**82.4**	71.8	52.6	86.2	38.9	41.2	15.5	60.3	18.1	12.1	2.4
Ours (3D ECA)	97.8	79.4	71.9	53.3	88.2	41.4	40.6	16.2	57.1	19.4	14.8	2.9
NeRF-RPN (anchor-free)	96.3	69.9	85.2	59.9	89.2	42.9	55.5	18.4	66.7	27.3	30.9	11.5
Ours (3D SE-CA)	97.8	75.7	86.4	64.5	89.7	43.8	56.9	**23.5**	**71.4**	**31.8**	33.4	**20.1**
Ours (3D ECA)	97.1	77.2	**86.9**	65.4	**91.6**	**44.3**	58.8	23.4	70.2	30.2	**35.4**	18.4

Fig. 4. Qualitative results of 3D object detection on the 3D-FRONT dataset. Compared to the baseline method NeRF-RPN, our method demonstrates superior performance in accurately predicting the 3D bounding boxes in 3D space.

reconstructed by NeRF in 3D space (see last row). 3D geometry in NeRF plays a significant role in region proposal tasks and poor NeRF reconstruction can seriously impair the prediction. However, our method can still detect them accurately compared to the baseline method.

C. Ablation Studies

We conducted the ablation study to examine the effect of adding a channel attention module to the output of each bottom-up feature map from the backbone network before passing them to corresponding top-down feature maps. To determine the optimal placement, we tested the module between various combinations of feature maps: $\{C5\}$ and $\{P5\}$, $\{C4, C5\}$ and $\{P4, P5\}$, $\{C3, C4, C5\}$ and $\{P3, P4, P5\}$, and finally among all feature maps, as shown in Table II. From the experiment results, it is evident that the best performance is attained when the channel attention module is located between all feature map levels $\{C2, C3, C4, C5\}$ and $\{P2, P3, P4, P5\}$. This configuration allows feature maps $\{P2, P3, P4, P5\}$ to effectively acquire channel-wise global information, thus enhancing detection accuracy.

TABLE II

ABLATION STUDY OF ADDING CHANNEL ATTENTION MODULE AT DIFFERENT LEVELS OF FEATURES BETWEEN THE 3D VGG19 BACKBONE AND 3D FPN ON 3D-FRONT DATASET. LEVELS 1, 2, AND 3 REPRESENT THE NUMBER OF LAYER COMBINATIONS WE UTILIZED.

Methods	3D-FRONT			
	R@0.25	R@0.5	AP@0.25	AP@0.5
3D ECA LEVEL1	97.0	73.5	83.6	60.0
3D ECA LEVEL2	97.0	74.3	86.0	63.6
3D ECA LEVEL3	96.3	75.7	86.0	63.8
Ours	**97.1**	**77.2**	**86.9**	**65.4**

V. CONCLUSIONS

In this paper, we proposed effective attention-based feature calibration mechanisms to enhance the performance of NeRF-based 3D detection from multi-view RGB images. Specifically, we investigated 3D SE-CA and 3D ECA modules, to incorporate global information into 3D feature maps extracted from the backbone network to enhance feature representation. The downstream detection network based on the enhanced features, with more discriminative power, achieves better performance in the challenging domain of 3D space. Our experimental results demonstrate that our method outperforms the baseline NeRF-RPN on all three datasets: Hypersim by 8.6 AP@0.5, 3D FRONT by 5.5 AP@0.5, and ScanNet by 5.1 AP@0.5, with an anchor-free RPN head. These results not only illustrate the importance of channel attention mechanisms in 3D object detection tasks, but also provide valuable insights into the critical factors to achieve optimal performance in this area.

REFERENCES

[1] J. Sun, Y. Xu, M. Ding, *et al.*, "Nerf-loc: Transformer-based object localization within neural radiance fields," *IEEE Robotics and Automation Letters*, 2023.

[2] J. Mao, S. Shi, X. Wang, and H. Li, "3d object detection for autonomous driving: A comprehensive survey," *International Journal of Computer Vision*, vol. 131, no. 8, pp. 1909–1963, 2023.

[3] M. Martini, F. Solari, and M. Chessa, "Obstacle avoidance and interaction in extended reality: An approach based on 3d object detection," in *ICIAP*, 2023.

[4] W. Lee, N. Park, and W. Woo, "Depth-assisted real-time 3d object detection for augmented reality," in *ICAT*, vol. 11, 2011, pp. 126–132.

[5] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *CVPR*, 2018.

[6] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *CVPR*, 2019.

[7] C. R. Qi, O. Litany, K. He, and L. J. Guibas, "Deep hough voting for 3d object detection in point clouds," in *ICCV*, 2019.

[8] C. R. Qi, X. Chen, O. Litany, and L. J. Guibas, "Imvotenet: Boosting 3d object detection in point clouds with image votes," in *CVPR*, 2020.

[9] X. Shen and I. Stamos, "Frustum voxnet for 3d object detection from rgb-d or depth images," in *WACV*, 2020.

[10] T. Wang, J. Pang, and D. Lin, "Monocular 3d object detection with depth from motion," in *ECCV*, 2022.

[11] G. Brazil and X. Liu, "M3d-rpn: Monocular 3d region proposal network for object detection," in *ICCV*, 2019.

[12] D. Rukhovich, A. Vorontsova, and A. Konushin, "Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection," in *WACV*, 2022.

[13] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *ECCV*, 2020.

[14] B. Hu, J. Huang, Y. Liu, Y.-W. Tai, and C.-K. Tang, "Nerf-rpn: A general framework for object detection in nerfs," in *CVPR*, 2023.

[15] C. Xu, B. Wu, J. Hou, *et al.*, "Nerf-det: Learning geometry-aware volumetric representation for multi-view 3d object detection," in *CVPR*, 2023.

[16] C. Huang, Y. Hou, W. Ye, *et al.*, "Nerf-det++: Incorporating semantic cues and perspective-aware depth supervision for indoor multi-view 3d detection," *arXiv preprint arXiv:2402.14464*, 2024.

[17] C. Huang, X. Li, S. Zhang, L. Cao, and R. Ji, "Nerf-dets: Enhancing multi-view 3d object detection with sampling-adaptive network of continuous nerf-based representation," *arXiv preprint arXiv:2404.13921*, 2024.

[18] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *CVPR*, 2020.

[19] D. Rukhovich, A. Vorontsova, and A. Konushin, "Fcaf3d: Fully convolutional anchor-free 3d object detection," in *ECCV*, 2022.

[20] J. Gwak, C. Choy, and S. Savarese, "Generative sparse detection networks for 3d single-shot object detection," in *ECCV*, 2020.

[21] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering.," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[22] K.-C. Chan, J. Xiao, H. L. Goshu, and K.-m. Lam, "Point cloud densification for 3d gaussian splatting from sparse input views," in *ACM Multimedia 2024*.

[23] Z. Murez, T. Van As, J. Bartolozzi, A. Sinha, V. Badrinarayanan, and A. Rabinovich, "Atlas: End-to-end 3d scene reconstruction from posed images," in *ECCV*, 2020.

[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018.

[25] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *ECCV*, 2018.

[26] J. Xiao, Z. Lyu, C. Zhang, Y. Ju, C. Shui, and K.-M. Lam, "Towards progressive multi-frequency representation for image warping," in *CVPR*, 2024.

[27] J. Xiao, Q. Ye, T. Liu, C. Zhang, and K.-M. Lam, "Deep progressive feature aggregation network for multi-frame high dynamic range imaging," *Neurocomputing*, vol. 594, p. 127 804, 2024.

[28] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017.

[29] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010.

[30] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[31] R. Girshick, "Fast r-cnn," in *ICCV*, 2015.

[32] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *ICCV*, 2019.

[33] D. Zhou, J. Fang, X. Song, *et al.*, "Iou loss for 2d/3d object detection," in *3DV*, 2019.

[34] M. Roberts, J. Ramapuram, A. Ranjan, *et al.*, "Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding," in *ICCV*, 2021.

[35] H. Fu, B. Cai, L. Gao, *et al.*, "3d-front: 3d furnished rooms with layouts and semantics," in *ICCV*, 2021.

[36] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *CVPR*, 2017.

[37] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM transactions on graphics (TOG)*, vol. 41, no. 4, pp. 1–15, 2022.

[38] B. Roessle, J. T. Barron, B. Mildenhall, P. P. Srinivasan, and M. Nießner, "Dense depth priors for neural radiance fields from sparse input views," in *CVPR*, 2022.

[39] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.