DyFusion: Cross-Attention 3D Object Detection with Dynamic Fusion

Jiangfeng Bi, Haiyue Wei, Guoxin Zhang, Kuihe Yang, and Ziying Song

***Abstract*—In the realm of autonomous driving, LiDAR and camera sensors play an indispensable role, furnishing pivotal observational data for the critical task of precise 3D object detection. Existing fusion algorithms effectively utilize the complementary data from both sensors. However, these methods typically concatenate the raw point cloud data and pixel-level image features, unfortunately, a process that introduces errors and results in the loss of critical information embedded in each modality. To mitigate the problem of lost feature information, this paper proposes a Cross-Attention Dynamic Fusion (CADF) strategy that dynamically fuses the two heterogeneous data sources. In addition, we acknowledge the issue of insufficient data augmentation for these two diverse modalities. To combat this, we propose a Synchronous Data Augmentation (SDA) strategy designed to enhance training efficiency. We have tested our method using the KITTI and nuScenes datasets, and the results have been promising. Remarkably, our top-performing model attained an 82.52% mAP on the KITTI test benchmark, outperforming other state-of-the-art methods.**

Link to graphical and video abstracts, and to code: https://latamt.ieeer9.org/index.php/transactions/article/view/8434

***Index Terms*—Cross-Attention Dynamic Fusion, Synchronous Data Augmentation, 3D object detection**

I. Introduction

Against the backdrop of thriving autonomous driving advances, 3D object detection has arisen as an imperative task to equip unmanned vehicles with precise environmental cognition [1]. Pioneers in 3D object detection have carried out significant research, demonstrating excellent performance on public datasets such as KITTI [2] and nuScenes [3]. As an exemplar pioneering work, Qi et al. [4] devised PointNet, an innovative deep neural architecture that directly learns global features from point cloud data. Zhou et al. [5] proposed VoxelNet with a Voxel Feature Encoder, which transforms raw point clouds into voxel-wise features containing spatial and physical information. These pioneering methods have laid a strong foundation and provided inspiration for subsequent research in the realm of 3D object detection.

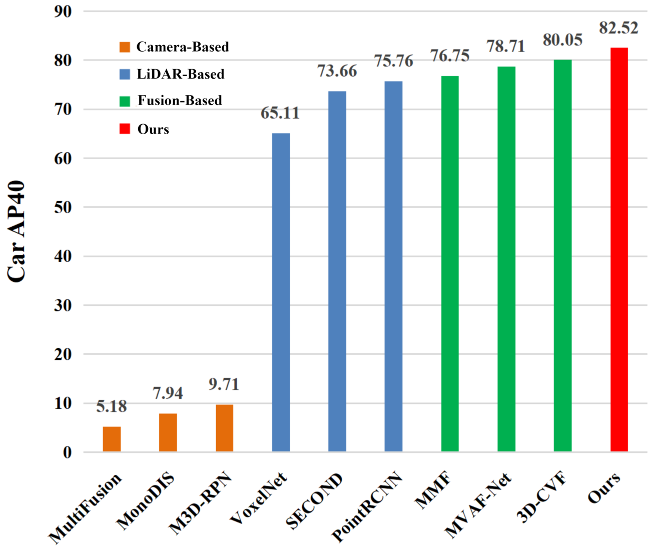


Fig. 1. Quantitative Analysis of Camera, LiDAR, Fusion and Proposed Methods.

TABLE I
Comparison of Advantages and Disadvantages of Camera and LiDAR

Sensor	Advantages	Disadvantages
Camera	High resolution, color information.	Sensitive to lighting and weather, difficult to handle reflective surfaces.
LiDAR	Distance information, no need for lighting.	Lower resolution, difficulty in recognizing color and texture.

Within the realm of autonomous driving, LiDAR and camera serve as key sensors, providing rich data for 3D object detection [1]. These two sensors have highly complementary output data, with their respective advantages and disadvantages summarized in Table I. Fusion-based algorithms tested on the KITTI dataset demonstrate better detection performance compared to using only camera or LiDAR alone, as shown in Fig. 1. Consequently, fusion-based technology has attracted significant research interest. Frustum PointNets [6], an end-to-end 3D object detection method devised by Chen et al., integrates 2D object detection outputs with point cloud data. Building on the foundation of MV3D [7], Ku et al. [8] introduced AVOD, a fusion-based approach utilizing both image and point cloud features to achieve more accurate 3D object detection.

Point cloud data contains the coordinate information of 3D spatial points in a scene, providing high precision and reliability. Image data provides high-resolution information

This work was supported by the Natural Science Foundation of Hebei Province, Grant No. F2013208105. (Corresponding author: Ziying Song.)

Jiangfeng Bi, Haiyue Wei, Guoxin Zhang, and Kuihe Yang are with the School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang 050018, China (e-mail: bjf19981227@163.com; ezio59624@gmail.com; zhangguoxincs@gmail.com; ykh@hebust.edu.cn).

Ziying Song is with the School of Computer and Information Technology, Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing, China. (e-mail: 22110110@bjtu.edu.cn).

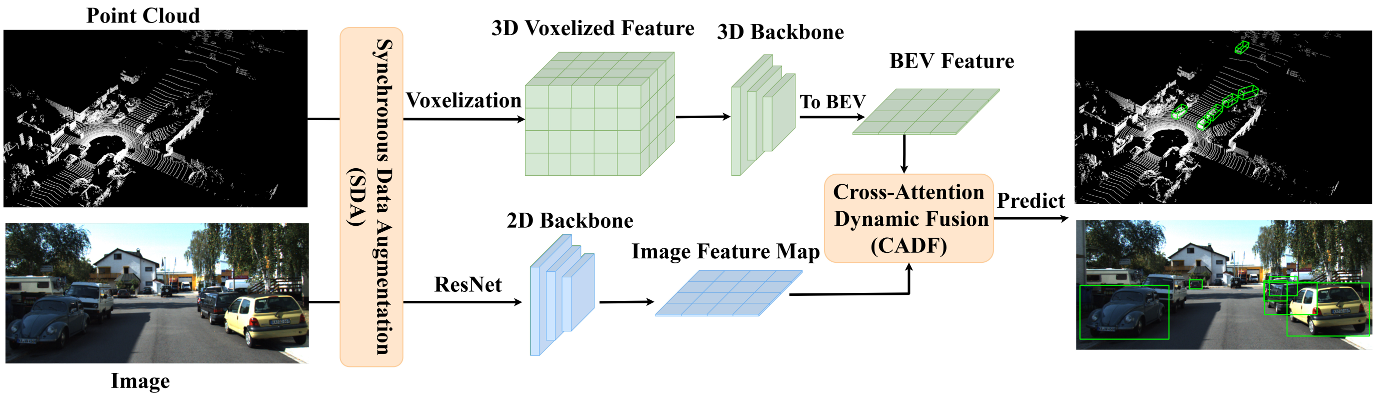


Fig. 2. The overall design of the DyFusion approach. It consists of two core components: the CADF (Section II-A) and the SDA (Section II-B). The CADF utilizes cross-attention to fuse point cloud and image features. The SDA employs a simultaneous enhancement strategy to improve model robustness.

such as object color, texture, and shape with visual richness. Although pure camera-based detection can capture image information, the lack of depth cues hinders accurate 3D localization, degrading accuracy. Moreover, pure point cloud methods struggle with sparsity and disorder, risking missed distant or occluded targets and thus detection errors. By fusing these two heterogeneous data sources, the accuracy of object detection can be significantly improved [1]. Previous research has primarily used element-wise addition or concatenation for fusion, but these strong fusion approaches risk information loss of semantic cues from images and depth patterns from point clouds, potentially impairing model performance. Therefore, effectively integrating these two heterogeneous data types into a highly accurate and coherent representation remains hard to solve.

To tackle this issue, we propose the cross-attention dynamic fusion (CADF) strategy, which utilizes the cross-attention mechanism to dynamically combine cross-modal features. The model is calibrated by dynamic weighting, reducing the information loss and decreasing error. In addition, we introduce the synchronized data augmentation (SDA) strategy for heterogeneous data to enhance the robustness of the model and the problem of insufficient data augmentation for heterogeneous modalities. Through the synchronous enhancement of both point cloud features and image data, our model is able to learn more universal informational representations, thus elevating its generalization capabilities by acquiring more transferable embeddings.

Our contributions are summarized into the following three aspects:

- We devise a **Cross-Attention Dynamic Fusion strategy** that effectively integrates point cloud and image information to improve object detection performance.
- We introduce a **Synchronous Data Augmentation strategy** for heterogeneous data that enhances the accuracy and robustness of the model's feature extraction.
- Our proposed approaches exhibit compelling performance in rigorous empirical evaluations on the authoritative KITTI and nuScenes benchmarks, which verifies their efficacy and validity in real-world autonomous driving scenarios.

The structure of this paper is arranged as follows: Section II describes our proposed method in detail. Section III introduces the experimental setup, evaluation metrics, and results in detail. Section IV concludes the paper and outlooks future work.

II. Method

This section details our proposed DyFusion model. As shown in Fig. 2, DyFusion fuses point clouds and RGB images via cross-attention to enhance 3D object detection.

A. Cross-Attention Dynamic Fusion

Traditional multimodal methods in computer vision often rely on simple fusion strategies like arithmetic operations and concatenation to integrate image and point cloud features. However, these strategies have two potential issues that can reduce fusion effectiveness. First, ascribed to the dispersed nature of point cloud inputs, directly fusing dense image features onto point clouds can result in lost semantic information from the image. Second, this approach may introduce data redundancy, hindering important information aggregation and introducing noise.

In order to tackle these challenges, we propose a novel Cross-Attention Dynamic Fusion (CADF) strategy for dynamically fusing heterogeneous data sources. The CADF framework diagram is shown in Fig. 3. Unlike traditional approaches, CADF dynamically fuses features with dynamic weighting to allow each point cloud to capture better and fuse image features, thus more effectively utilizing the global semantic information in the image. In the succeeding sections, we elaborate on the proposed CADF methodology by delineating its architectural schematics.

1) Image Feature Extraction: To extract global feature representations from raw image inputs, the proposed CADF incorporates a ResNet-50 architecture as the backbone module for 2D feature learning. ResNet-50, which is widely adopted in computer vision tasks such as Multi-Foggy Images [9]

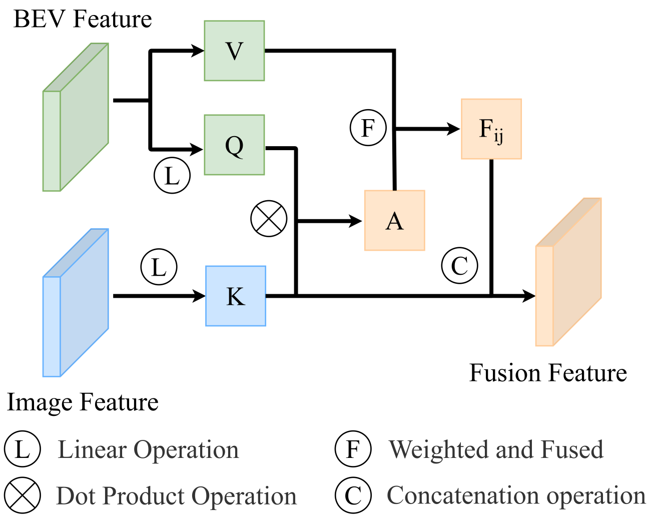


Fig. 3. The framework of CADF. First, the BEV and image features are transformed linearly into Q and K matrices, respectively. These are multiplied to obtain an attention matrix A, which is then normalized. The resulting matrix is used to weigh and fuse the BEV and image features. The fused features are obtained by concatenating the resulting matrix with the original image features.

and ConvGRU-CNN [10]. In our approach, ResNet-50 extracts global feature maps from raw images, generating high-dimensional feature representations $h_x \in \mathbb{R}^d$ on a per-image basis. Where d represents the dimension of the feature vector, which can be expressed as:

$$h_x = f_{\theta_x}(x) \tag{1}$$

Where f_{θ_x} represents the ResNet-50 network, where θ_x denotes the parameters of the network.

2) Voxel Feature Encoding: The CADF strategy uses Voxel Feature Encoding (VFE) [5] to process point cloud data. The point cloud data $P = p_{i}{}_{i=1}^{N} \in \mathbb{R}^{N\times 3}$ is processed through the VFE layer to obtain the feature map $v_j \in \mathbb{R}^d$ for each voxel, where d denotes the dimension of the feature vector and j denotes the index of the voxel. It can be expressed as:

$$v_j = f_{\theta_p}(P_j) \tag{2}$$

Where f_{θ_p} represents the VFE layer, where θ_p denotes the parameters of the VFE layer, and P_j represents the point cloud data in the j-th voxel.

3) BEV Feature Extraction: Subsequently, the voxel feature representations are projected onto the BEV (Bird's Eye View) plane through orthogonal mapping, acquiring the positional coordinates and associated feature vectors of the point cloud data in the BEV perspective. The feature vectors of all voxels are aggregated according to certain rules to obtain a voxel feature vector $V \in \mathbb{R}^{H\times W\times d}$, where H and W denote the height and width of the BEV plane. It can be expressed as:

$$V_{i,j,k} = g(v_{l,m,n}) \tag{3}$$

where i, j denotes the position coordinates in the BEV plane, k denotes the dimensionality of the feature vector, l, m, n denotes the index of the voxel, and g denotes the aggregation function, which can be pooling, averaging, or maximization functions.

4) Feature Dynamic Fusion: We use the Cross-Attention mechanism to fuse BEV features with image features to achieve feature fusion. Cross-attention mechanisms can interact and integrate information from different feature spaces to produce a more expressive feature representation. The CADF module allows each point cloud to blend the image characteristics more optimally, better utilize the global semantic information of the image, and better handle the input features of the two heterogeneous modalities.

First, the BEV and image features are separately transformed by linear operations to obtain two new feature maps, $Q \in \mathbb{R}^{H\times W\times d}$ and $K \in \mathbb{R}^{H\times W\times d}$, where d represents the dimensionality of the feature vectors. This can be expressed as follows:

$$Q = W_q V, \;\; K = W_k h_x \tag{4}$$

Where V denotes the BEV feature matrix, h_x denotes the image feature vector, and W_q and W_k are the linear transformation matrices that can be learned.

Next, the dot product operation of Q and K yields an attention matrix $A \in \mathbb{R}^{H\times W\times H\times W}$, which can be expressed as

$$A_{i,j,k,l} = \frac{1}{\sqrt{d}} Q_{i,j,:} \cdot K^T_{k,l,:} \tag{5}$$

where i, j and k, l denote the position coordinates in the attention matrix, $:$ denotes all the feature vectors, and $\sqrt{d}$ is used to scale the dot product results so as to avoid the problem of exploding or vanishing gradients.

Then, the attention matrix A is subjected to softmax operation to obtain a normalized attention matrix $P \in \mathbb{R}^{H\times W\times H\times W}$, which can be expressed as

$$P_{i,j,k,l} = \frac{\exp(A_{i,j,k,l})}{\sum_{m,n} \exp(A_{i,j,m,n})} \tag{6}$$

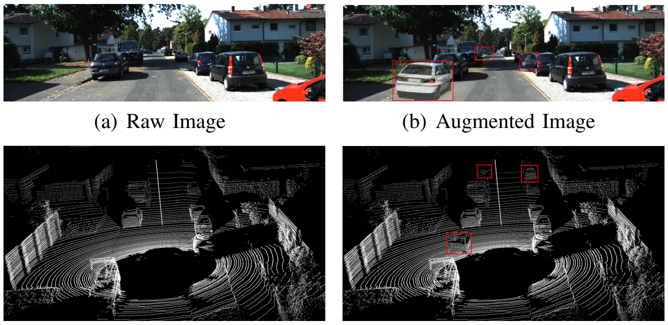
Next, the BEV feature matrix V is weighted and fused using P to obtain a fused feature matrix $F \in \mathbb{R}^{H\times W\times d}$, which is formulated as

$$F_{i,j,:} = \sum_{k,l} P_{i,j,k,l} V_{k,l,:} \tag{7}$$

Finally, the fused feature matrix F and the image feature vector h_x are concated together to obtain a fused feature vector $h_f \in \mathbb{R}^{2d}$, which can be expressed as

$$h_f = [F_{i,j,:}, h_x] \tag{8}$$

Notably, the attention matrix A in the cross-attention module quantifies feature similarity between the BEV representations and image embeddings, determining the weight assignment of features at different locations during fusion. To achieve this, softmax operations map the values in A to the range [0,1] for weighted fusion. The resulting fusion output h_f contains both BEV and image features, thus providing an enhanced representation of the spatial relationship and semantic information of point clouds and images.



(a) Raw Image (b) Augmented Image

(c) Raw Point (d) Augmented Point

Fig. 4. An illustration of Synchronous Data Augmentation.

B. Synchronous Data Augmentation

Pioneers have done substantial work on data augmentation to enhance the diversity and quantity of data in 3D object detection. However, most augmentation strategies are applied only to point cloud data, which increases sample size and avoids efficiency reduction caused by point cloud sparsity. Nevertheless, augmenting only point clouds ignores consistency and correspondence with image data, potentially leading to model overfitting.

To address this risk and improve model generalization, we propose a Synchronous Data Augmentation (SDA) strategy, visualized in Fig. 4. These samples were randomly selected from the KITTI dataset [2]. We perform synchronous augmentation on corresponding image data while augmenting point clouds, maintaining correspondence between modalities. This SDA strategy avoids inconsistency issues during training, thereby improving model accuracy and robustness. Additionally, SDA can simultaneously apply various transformations to point cloud and image data, increasing dataset diversity. This enables the model to better adapt to different scenarios and environments.

1) Point Cloud Data Enhancement: To eliminate noise and increase dataset diversity, we applied various data augmentation techniques to point clouds, drawing inspiration from VoxelNet [5] and SECOND [11]. We perform random rotation, scaling, translation, clipping, and subsampling, explained as follows. Point clouds underwent rotations randomly sampled from [-π/4, π/4]. We augmented data via random scaling in [0.95, 1.05] and translation in [-0.1, 0.1]. For random cropping, parameters were randomly chosen from [-0.1, 0.1]. Subsampling reduced each point cloud to 512 points. These augmentations significantly increased point cloud diversity, improving model robustness and generalization.

Randomly selecting and concatenating subsets of objects from the training dataset into each frame increases the number and diversity of objects, improving training speed and efficiency. Point cloud sparsity means frames may contain few objects, preventing networks from fully learning object features. Introducing more objects can thus improve model accuracy and robustness. Additionally, more objects enhance adaptability to varied scenes, improving generalization. Therefore, this augmentation method effectively enhances model performance.

2) Image Data Enhancement: We devise an innovative method to enhance the awareness of image data. We fuse depth data extracted from 3D object label annotations with the Mixup approach to achieve this goal. Specifically, we first align the raw image with the augmented point cloud, then execute partial cropping. The corresponding transparency is then applied to each cropped region according to its depth information, thus representing depth information in the image.

To generate the enhanced image, we first sort the target objects in order of their depth and crop the same area from the original image for each object. We then blend each cropped region with the corresponding target image using the blending ratio in the Mixup method. This ensures that the enhanced image maintains continuity with the original image. Additionally, we attenuate the transparency of each object's region in the enhanced image according to its depth order. This guarantees consistency between the enhanced image and the original image.

By incorporating depth information into the image data via our proposed method, our model can learn from both the point cloud and augmented image representations, thereby enhancing performance and robustness. In summary, our method is an effective data augmentation technique that enhances the depth-awareness of image data.

III. Experiments

A. Dataset and Metrics

1) KITTI Dataset: The KITTI [2] dataset pioneers as a benchmark tailored for autonomous driving scenarios. Captured using cameras and LiDARs mounted on a driving vehicle, it portrays complex urban and highway environments. Pixel-level annotations further enrich the dataset, providing ground truths for salient vision tasks including object detection, depth completion and semantic segmentation. The imagery comprises high-resolution stereo pairs, while LiDAR sweeps depict detailed 3D spaces. Owing to its comprehensive coverage of driving perception challenges, KITTI continues to catalyze innovations in autonomous vehicle research.

2) nuScenes Dataset: Released by Motional LLC, the nuScenes [3] dataset pioneers as a large-scale benchmark tailored for autonomous driving. Captured in Boston and Singapore by a fleet of AVs, it comprises 1.5 million images and multi-sensor data from cameras, LiDARs and radars. The 360-degree sensor suite provides comprehensive coverage of driving scenes. Key attributes include 1600×900 images at $12Hz$, dense 3D point clouds and radars detecting beyond 200 meters. nuScenes also incorporates meticulous semantic annotations like 3D boxes and tracking labels.

3) Metrics: The mAP (mean Average Precision) calculates the average detection accuracy for each class as :

$$mAP = \frac{1}{C} \times \sum_{c \in C} (AP_c) \tag{9}$$

where C represents the total number of categories. The NDS (nuScenes detection score) is weighted sum of mAP and er-

TABLE II
QUANTITATIVE ANALYSIS OF DYFUSION ON KITTI TEST SET FOR CAR CATEGORY BASED ON 40 RECALL POINTS AP. "MOD." REPRESENTS MODERATE. "-" MEANS NOT MENTIONED. "C" REPRESENTS CAMERA. "L" REPRESENTS LIDAR. "L&R" REPRESENTS LIDAR AND CAMERA FUSION

Method	Modality	AP_{3D} (%)			AP_{BEV}(%)		
		Easy	Mod.	Hard	Easy	Mod.	Hard
MonoDIS [18]	C	10.37	7.94	6.40	17.34	13.19	11.12
M3D-RPN [19]	C	14.76	9.71	7.42	-	-	-
VoxelNet [5]	L	77.47	65.11	57.73	89.35	79.26	77.39
SECOND [11]	L	84.65	73.66	68.71	88.07	79.37	77.50
MVAF-Net [15]	L&R	87.87	78.71	75.48	91.95	87.73	85.00
3D-CVF [16]	L&R	89.20	80.05	73.11	**93.52**	89.56	82.45
CLOCs [17]	L&R	88.94	80.67	77.15	93.05	**89.80**	86.57
Fast-CLOCs [20]	L&R	89.11	80.34	76.98	93.02	89.49	86.39
Baseline [12]	L	90.90	81.62	77.06	-	-	-
DyFusion (Ours)	L&R	**90.96**	**82.52**	**77.91**	93.02	89.20	**86.57**

rors, leading to a more comprehensive description of detection performance, as:

$$\mathrm{NDS} = \frac{1}{10}\left[5\ \mathrm{mAP} + \sum_{\mathrm{mTP}\in\mathrm{TP}} (1 - \min(1, \mathrm{mTP}))\right] \quad (10)$$

where mTP(mean True Positive) denotes the average distance threshold for each category.

B. Implementation Details

In the experiment, we chose advanced methods Voxel R-CNN [12] and CenterPoint [13] for the point cloud branch and validated our module on the KITTI and nuScenes datasets. For the image branch, considering the trade-off between speed and accuracy, we adopt ResNet-50 [14] to extract image features following previous works [15]–[17]. In the experiment, the voxel size was configured as (0.1m, 0.1m, 0.1m) along the x, y, and z axes. Our DyFusion training framework uses a mixed optimizer for end-to-end optimization. The detailed configuration of the experiment is as follows: the experiment is trained on 8 RTX 3090 GPUs; a total of 80 epochs are trained; the batch size is set to be 8; and the experiment is conducted using Python 3.8, PyTorch 1.10 and CUDA 11.3. The network was trained for approximately nine hours. We used OpenPCDet as the codebase, and if no parameter settings were specified, the default settings were applied.

C. Results on KITTI Dataset

To examine the proposed architecture, experiments were conducted on the KITTI dataset [2] and reported the average accuracy (AP40). Our proposed DyFusion achieves significant results. Specifically, DyFusion improves the accuracy by 0.9, 1.79, and 1.86 in three different classes compared to baseline Voxel R-CNN [12]. We also compared some advanced 3D object detection methods, and Table II presents the results of our method and other latest techniques evaluated on the KITTI test data. We also performed validation experiments, and Table III shows our quantitative results on the KITTI validation set.

TABLE III
QUANTITATIVE ANALYSIS OF DYFUSION ON KITTI VALIDATION SET FOR CAR CATEGORY BASED ON 40 RECALL POINTS AP

Method	AP_{3D} (%)			AP_{BEV} (%)		
	Easy	Mod.	Hard	Easy	Mod.	Hard
VoxelNet [5]	81.97	65.46	62.85	-	-	-
F-PointNets [21]	83.76	70.92	63.65	-	-	-
SECOND [11]	87.43	76.48	69.10	-	-	-
Point-GNN [22]	87.89	78.34	77.38	89.82	88.31	87.16
PointRCNN [23]	88.88	78.63	77.38	-	-	-
F-ConvNet [24]	89.02	78.80	77.09	90.23	88.79	86.84
MAFF-Net [25]	88.88	79.37	74.68	93.23	89.31	86.61
EPNet [26]	92.28	82.59	80.14	95.51	91.47	**91.16**
Baseline [12]	92.39	84.55	82.03	95.68	90.77	88.46
DyFusion(Ours)	**93.29**	**86.34**	**83.89**	**96.57**	**92.08**	90.24

D. Results on nuScenes Dataset

In order to assess the efficacy and viability of our approach, we conducted supplementary experiments on the nuScenes [3] dataset. The results in Table IV show that DyFusion achieved 62.8 mAP and 68.5 NDS on the nuScenes dataset, which is a significant improvement of 4.8 mAP and 3.0 NDS compared to the baseline CenterPoint [13]. DyFusion also outperformed many advanced methods in terms of detection accuracy. We also provide detailed results for each object class and performance on the test leaderboard in Table IV.

E. Ablation Studies

To demonstrate the improvement of detection accuracy by the CADF strategy and SDA strategy in DyFusion, we tested both strategies on the baseline. We benchmarked the approaches on the KITTI val data and exhibited the AP results in Table V. After adding our CADF strategy to the model, the Car AP3D increased from 84.55 to 85.98, indicating that the dynamic fusion strategy for images and point clouds plays a crucial role in our network framework. Furthermore, the SDA strategy that we introduced further improved the training accuracy of the model by 0.36 mAP. The improvement in model accuracy demonstrates that our method can better utilize the depth information from point clouds and semantic information from images for 3D object detection. Additionally, synchronized data augmentation enhances the robustness of the model and optimizes the entire network structure.

F. Results and Discussion

In this section, we experimentally validate the proposed DyFusion model, which is trained and evaluated using KITTI and nuScenes datasets. The experimental results show that the DyFusion model significantly improves the accuracy in 3D object detection tasks. However, we should also note that the

TABLE IV
COMPARISONS WITH PREVIOUS METHODS ON NUSCENES TEST SET

Method	mAP	NDS	Car	Truck	C.V.	Bus	Trailer	Barrier	Motor.	Bike	Ped.	T.C.
SECOND [11]	31.6	46.8	-	-	-	-	-	-	-	-	-	-
PointPainting [27]	46.4	58.1	77.9	35.8	15.8	36.2	37.3	60.2	41.5	24.1	73.3	62.4
HotSpotNet [28]	59.3	66.0	83.1	50.9	56.4	53.3	23.0	81.3	63.5	36.6	73.0	71.6
CenterPoint [13]	58.0	65.5	84.6	51.0	**60.2**	53.2	17.5	**83.4**	53.7	28.7	76.7	70.9
DyFusion(Ours)	**62.8**	**68.5**	**86.0**	**58.8**	21.8	**72.0**	39.5	66.0	**68.9**	**51.9**	**86.3**	**76.9**

TABLE V
PERFORMANCE OF PROPOSED METHOD WITH DIFFERENT STRATEGIES ON KITTI VAL SET

CADF	SDA	AP_{3D} (%)			AP_{BEV} (%))		
		Easy	Mod.	Hard	Easy	Mod.	Hard
		92.39	84.55	82.03	95.68	90.77	88.46
✓		93.12	85.98	83.52	96.13	91.89	89.97
	✓	92.85	85.22	82.87	90.04	91.26	88.74
✓	✓	**93.29**	**86.34**	**83.89**	**96.57**	**92.08**	**90.24**

experiments have some limitations. First, the complexity of the model and the computational resources limit the experimental scale and depth, which still puts us at a disadvantage compared to some state-of-the-art models (e.g., TransFusion [29], BEV-Fusion [30], and UVTR [31], etc.). Secondly, the prediction speed is slow due to the large computational volume and many parameters. Based on the experimental results and limitations, we propose the following recommendations to guide future research. First, expand the size and diversity of the training dataset to improve the generalization ability and robustness of the network. Second, further optimize the network architecture and hyperparameter settings to find lightweight neural network architectures, such as Densely Feature selection Convolutional neural Network – Hyper Parameter tuning [32].

IV. CONCLUSION

In this paper, we proposed DyFusion, a novel image-point cloud fusion method. Our method introduces CADF, an innovative approach that leverages cross-attention to effectively integrate image and point cloud features. This integration not only mitigates errors but also addresses the challenge of feature loss. We also designed the SDA strategy to address the problem of insufficient data augmentation, which helps introduce more variations into the training data thereby enhancing the robustness and generalization capability of the model. Extensive experiments on KITTI [2] and nuScenes [3] demonstrate DyFusion's effectiveness and superiority over other advanced 3D detection methods. We hope our proposed fusion method provides a new perspective to advance 3D object detection. In future work, we can further explore multimodal network lightweighting for faster detection and real-time applications in autonomous driving. This requires us to investigate more efficient feature extraction and fusion mechanisms, as well as exploring techniques such as model pruning and quantization to reduce model parameters and computation.

REFERENCES

[1] L. Wang, X. Zhang, Z. Song, J. Bi, G. Zhang, H. Wei, L. Tang, L. Yang, J. Li, C. Jia *et al.*, "Multi-modal 3d object detection in autonomous driving: A survey and taxonomy," *IEEE Transactions on Intelligent Vehicles*, 2023.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 918–927.

[7] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 1907–1915.

[8] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, "Joint 3d proposal generation and object detection from view aggregation," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1–8.

[9] Z. H. Arif, M. A. Mahmoud, K. H. Abdulkareem, S. Kadry, M. A. Mohammed, M. N. Al-Mhiqani, A. S. Al-Waisy, and J. Nedoma, "Adaptive deep learning detection model for multi-foggy images." *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 7, no. 7, 2022.

[10] M. Q. Gandapur and E. Verdú, "Convgru-cnn: Spatiotemporal deep learning for real-world anomaly detection in video surveillance system," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 4, pp. 88–95, 2023.

[11] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.

[12] J. Deng, S. Shi, P. Li, W. Zhou, Y. Zhang, and H. Li, "Voxel r-cnn: Towards high performance voxel-based 3d object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1201–1209.

[13] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3d object detection and tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 784–11 793.

[14] 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.

[15] G. Wang, B. Tian, Y. Zhang, L. Chen, D. Cao, and J. Wu, "Multi-view adaptive fusion network for 3d object detection," *arXiv preprint arXiv:2011.00652*, 2020.

[16] J. H. Yoo, Y. Kim, J. Kim, and J. W. Choi, "3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*. Springer, 2020, pp. 720–736.

[17] S. Pang, D. Morris, and H. Radha, "Clocs: Camera-lidar object candidates fusion for 3d object detection," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 386–10 393.

[18] A. Simonelli, S. R. Bulo, L. Porzi, M. López-Antequera, and P. Kontschieder, "Disentangling monocular 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1991–1999.

[19] G. Brazil and X. Liu, "M3d-rpn: Monocular 3d region proposal network for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9287–9296.

[20] S. Pang, D. Morris, and H. Radha, "Fast-clocs: Fast camera-lidar object candidates fusion for 3d object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 187–196.

[21] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 918–927.

[22] W. Shi and R. Rajkumar, "Point-gnn: Graph neural network for 3d object detection in a point cloud," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1711–1719.

[23] S. Shi, X. Wang, and H. Li, "Pointrcnn: 3d object proposal generation and detection from point cloud," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 770–779.

[24] Z. Wang and K. Jia, "Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 1742–1749.

[25] Z. Zhang, Y. Shen, H. Li, X. Zhao, M. Yang, W. Tan, S. Pu, and H. Mao, "Maff-net: Filter false positive for 3d vehicle detection with multi-modal adaptive feature fusion," in *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2022, pp. 369–376.

[26] T. Huang, Z. Liu, X. Chen, and X. Bai, "Epnet: Enhancing point features with image semantics for 3d object detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 35–52.

[27] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "Pointpainting: Sequential fusion for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4604–4612.

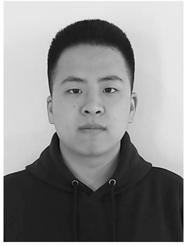
[28] Q. Chen, L. Sun, Z. Wang, K. Jia, and A. Yuille, "Object as hotspots: An anchor-free 3d object detection approach via firing of hotspots," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*. Springer, 2020, pp. 68–84.

[29] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "Transfusion: Robust lidar-camera fusion for 3d object detection with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1090–1099.

[30] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *2023 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2023, pp. 2774–2781.

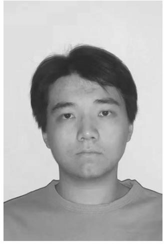
[31] Y. Li, Y. Chen, X. Qi, Z. Li, J. Sun, and J. Jia, "Unifying voxel-based representation with transformer for 3d object detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 18 442–18 455, 2022.

[32] M. Adimoolam, S. Mohan, J. A., and G. Srivastava, "A novel technique to detect and track multiple objects in dynamic video surveillance systems," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 4, pp. 112–120, 2022.



Jiangfeng Bi was born in Shijiazhuang, Hebei Province, China, in 1998. He received his B.S. degree from Hebei University of Science and Technology (China) in 2021. He is now a master's student majoring in Computer Science and Technology at the Hebei University of Science and Technology (China), mainly engaging in computer vision-related research during school.

Haiyue Wei was born in Shijiazhuang, Hebei Province, China, in 1997. He received his B.S. degree from the Hebei University of Science and Technology (China) in 2020. He is now a master's student majoring in Computer Science and Technology at the Hebei University of Science and Technology (China), mainly engaging in computer vision-related research.



Guoxin Zhang Guoxin Zhang, was born in 1998 in Xingtai, Hebei Province, China. In 2021, he received his Bachelor's degree. He is now studying for his master's degree at the Hebei University of Science and Technology (China). His research interests are in computer vision.

Kuihe Yang was born in 1966, in Handan, Hebei Province, China. He received the B.S. degree from Tianjin University (China) in 1988, the M.S. degree from University of Science and Technology Beijing (China) in 1997, and the Ph.D degree in computer application technology from Xidian University (China) in 2004. From 2005 to 2007,he was a Postdoctoral Fellow in Army Engineering University of PLA (China). He went to Manchester University (UK) for short-term training in 2011. Currently, He is professor and master tutor with Hebei University of Science and Technology (China). His research interests include database application technology, artificial intelligence and machine learning.

Ziying Song was born in Xingtai, Hebei Province, China in 1997. He received the B.S. degree from Hebei Normal University of Science and Technology (China) in 2019. He received a master's degree major in Hebei University of Science and Technology (China) in 2022. He is now a PhD student majoring in Computer Science and Technology at Beijing Jiaotong University (China), with a research focus on Computer Vision.