Soft and Hard Associations in Bi-modality Fusion Are All You Need

Chaokang Jiang[1]
jchaokang@gmail.com,

Abstract

Since different sensor signals are commonly heterogeneous, how to achieve feature-level fusion between different modalities in a most robust way remains a challenge in the field of robotics. In this paper, we propose a new fusion strategy, an adaptive bi-modality feature fusion module that combines both "soft fusion" and "hard fusion". This is a solution to the problem of non-robust fusion due to the poor data generated by one of the two sensor signals. Specifically, when a selected LiDAR point can be associated with a pixel of an image based on the sensor parameters, we use the multi-head attention mechanism to query the features in the LiDAR features and in the image features, respectively. We further design a point-wise "hard" association module to calculate the confidence scores of the two types of features and thus adaptively aggregate the associated features to this center point. Experiments on large-scale real-world dataset demonstrate that the proposed method outperforms the existing state-of-the-art methods. Compared to the baseline, hard fusion method and soft fusion method, our method improves by 51%, 30% and 4%, respectively.

1 Introduction

3D object detection task is receiving considerable attention in the field of intelligent robotics [Wang and Jia, 2019] and autonomous driving [Sagar, 2022], its main purpose is to estimate the localization, shape and specific semantics of an object from a given sensor signal. LiDAR and color cameras are commonly used sensors for 3D object detection tasks. Popular real-world datasets such as Waymo [Sun *et al.*, 2020], KITTI [Geiger *et al.*, 2013], nuScenes [Caesar *et al.*, 2020], etc. contain both sensors. However, both LiDAR signals and color images have inherent disadvantages. As shown in the upper left corner of Figure 1, the LiDAR signal can provide 3D geometric information of an object, but its sparseness leads to the loss of most information of small and distant objects. Color images contain dense pixels and provide rich texture information. But it is difficult to obtain the depth information of each pixel from a single image, which limits our ability to perceive the 3D distance information of an object. Therefore, bridging the disadvantage of a single sensor by fusing two sensors [Roy *et al.*, 2022] is a promising direction.

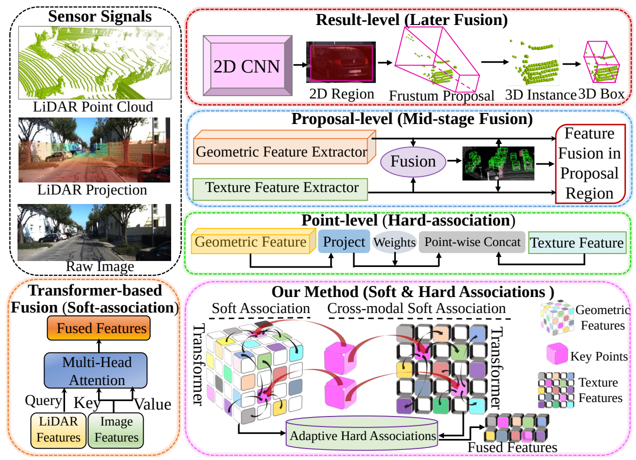

Figure 1: **Comparison of different multi-modality fusion solutions.** 3D sparse point cloud and 2D image fusion methods are classified into four categories: result-level, proposal-level, point-level, and Transformers-based fusion. We call the schemes that decorate LiDAR point clouds using image texture information in a point-by-point method as "Hard-association" at the point level. The schemes for LiDAR points to query and fuse global image features based on Transformers structure are called "Soft-association".

Many works have given different schemes on how to fuse LiDAR signals and images for improving autonomous driving perception performance more reasonably and efficiently. As shown in Figure 1, we classify deep learning-based 3D object detection methods into four categories: Result-level, Proposal-level, Point-level (hard-association) and Transformer-based Fusion (soft-association). The result-level fusion scheme, exemplified by Frustum-PointNets [Qi *et al.*, 2018], relies heavily on 2D detection of the image. Frustum-PointNets maps 2D image detection results into a 3D Frustum of view and instantiates the segmentation of 3D objects. This method [Qi *et al.*, 2018] has difficulty in estimating the 3D positions of small objects and heavily occluded objects because the final results are derived from the 2D detection results. In addition, MV3D [Chen *et al.*, 2017] and

AVOD [Ku *et al.*, 2018] directly fuse the two modality features in the region where the initial predicted proposal boxes are located. This proposal-level fusion solution inevitably involves the addition of background noise features to the fusion, which can cause incorrect feature representation in complex scenes. Next, there are also some works [Huang *et al.*, 2020; Jiang *et al.*, 2022] that perform a point-by-point correspondence between LiDAR points and pixels through the camera intrinsic and extrinsic. The point cloud features are augmented or decorated by the correspondence between LiDAR points and image pixels. The point-level feature fusion solutions represented by EPNet [Huang *et al.*, 2020] improves the performance of fusion-based 3D object detection networks up to one level. Further, the cross-modality data enhancement algorithm proposed by PointAugmenting [Wang *et al.*, 2021] further enhances the performance of point-level fusion methods. However, this point-level scheme is essentially a direct concatenation of features, which relies heavily on the calibration results between sensors. Recently, Transformer-based methods for solving soft correspondences of the multi-modality features have achieved in the best detection performance. Using the Transformer-based architecture, DeepFusion [Li *et al.*, 2022] fully considers feature alignment during fusion and physical alignment after data augmentation. TransFusion [Bai *et al.*, 2022] introduces simple and effective module for image-guided class-specific heatmap generation. These methods establish high-quality soft correlations between heterogeneous features, which effectively alleviate the loss of dense image features.

However, LiDAR and the camera do not always have high quality signal data at the same time [Bai *et al.*, 2022]. When some pixels are poorly textured representations affected by lighting, we should select better quality geometric features. When falling on the surface of certain objects with very sparse LiDAR points, we should select richer texture features. For more extreme cases, although the previous approach establishes a soft correspondence between high-quality bi-modality, this scheme still leads to a significant degradation of the network performance when the image data suffers from heavy contamination. Therefore, we consider that a more reasonable solution should consist of three elements: 1) First, each LiDAR point must have robust global geometric features. LiDAR features are the key to learn 3D object information.; 2) A LiDAR point features secondly be softly associated to rich image features; 3) A LiDAR point features should estimate confidence scores for its associated geometric and texture features, which makes the network aware that the current point is more supposed to focus on which high-quality features. Based on this inspiration, a bi-modality feature fusion module with both soft and hard components is designed in this paper. As shown in Figure 3, we propose a two-stage feature update method. The valid points in the initial prediction boxes of the first stage are used as queries. These queries focus on robust geometric and texture features from two modal features with Transformer-based soft feature association modules, respectively. Unlike previous methods that depend on soft fusion only or hard fusion only, our method aggregates the associated features into valid points in the prediction frame based on the computed feature confidence scores. Such soft and hard feature fusion methods effectively update the feature representation to improve the detection performance of the network. We demonstrate the effectiveness of the proposed method on a large-scale autonomous driving dataset, nuScenes [Caesar *et al.*, 2020]. The 32-beam LiDAR in the nuScenes dataset scans relatively sparse LiDAR points for small objects. Our fusion solution is further improving the performance of the network for small object detection. The main contributions of this work can be summarized in three main points:

- Harsh scenes and small objects with few signals cause no significant improvement in the accuracy of current 3D object detection networks. In this paper, a LiDAR-Camera fusion 3D detection framework is proposed and designed. The full potential of point cloud feature and image feature fusion is exploited, and a robust bi-modality fusion strategy is given especially for the poor signal on one side.
- A bi-modality feature fusion module with both hard and soft components is proposed, which guides the network to refine more accurate 3D positions and orientations of objects in the second stage. The rationality and effectiveness of this feature fusion scheme is demonstrated in this paper.
- The proposed method achieves state-of-the-art 3D detection performance on the nuScenes dataset, especially demonstrating powerful performance for small object detection with degraded image quality and objects with few LiDAR signals.

2 Related Works

2.1 Single-modality 3D Object Detection

The feature representation of the input signal is crucial for the 3D detection head to learn 3D object bounding box information. LiDAR point clouds are commonly used as input data for 3D object detection tasks. Many works use different forms of data to improve feature representation. PointNet [Qi *et al.*, 2017] learns the global spatial features of each point directly from the raw point cloud data. PointNet is the pioneer of deep learning of point clouds. F-PointNet [Qi *et al.*, 2018] performs 3D instance segmentation from the frustum of view to estimate the position of 3D objects based on PointNet. PointPillars [Lang *et al.*, 2019] extends point clouds from four-dimensional to nine-dimensional Pillars and extracts features from Pillars using PointNet. VoxelNet [Zhou and Tuzel, 2018] converts point cloud to voxels and proposes a Voxel Feature Encoding (VFE) module to learn point cloud features. CenterPoint [Yin *et al.*, 2021a] is based on VoxelNet using center points to represent objects, which simplifies the 3D object detection task. SECOND [Yan *et al.*, 2018] uses sparse convolution to effectively improve the disadvantage of more time-consuming 3D convolution.

2.2 Bi-modality 3D Object Detection

In recent years, there are many works focusing on point cloud and image fusion for 3D object detection. We roughly classify the different fusion solutions into four categories in Figure 1, where the first two categories of detection boxes [Qi

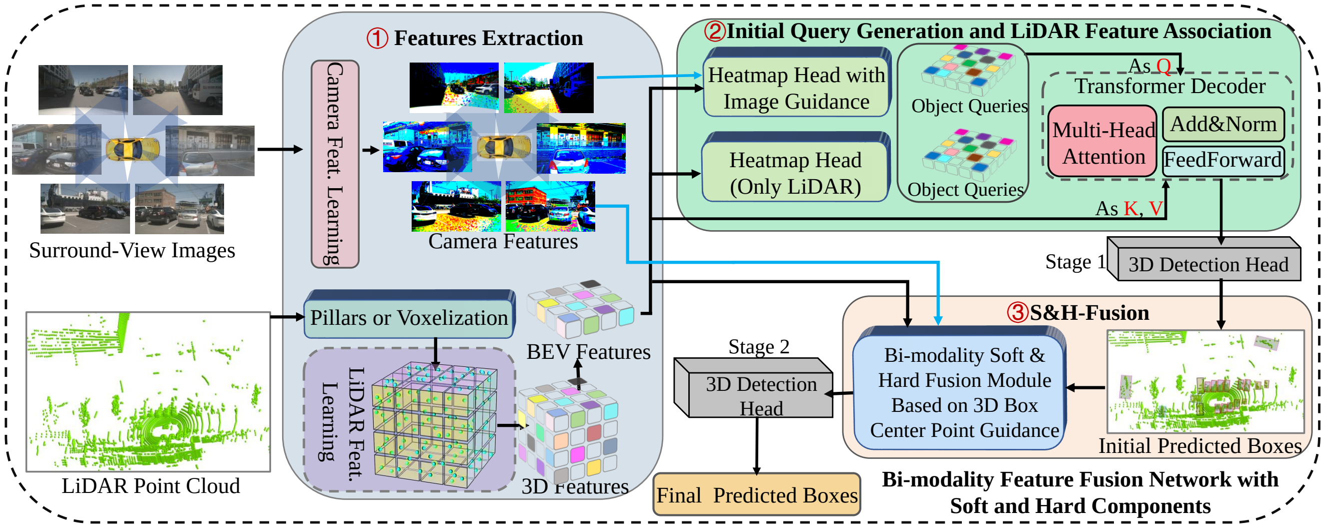

Figure 2: **The Main Network Structures of The Bi-modality Fusion Pipeline S&H-Fusion Proposed in This Paper**. The designed network is built on existing 2D image and 3D point cloud feature extraction networks to achieve subsequent feature fusion. The Bi-modality feature fusion stream consists of two key stages. Firstly, Object Queries are generated based on LiDAR BEV features and camera features. The initial object bounding box is estimated using Transformer Decoder and detection head. Secondly, the proposed soft and hard fusion strategy utilizes LiDAR features, image features and the initial 3D object box to further update the features, which is more focused on the objects detected in the first stage. Finally the final 3D object box is estimated using the 3D detection head.

et al., 2018; Shin *et al.*, 2019] and proposal boxes [Chen *et al.*, 2017; Ku *et al.*, 2018] based solutions are more shallow feature fusion, and these methods suffer from severe performance degradation in more complex environments. Later, some methods [Huang *et al.*, 2020; Vora *et al.*, 2020] obtain a one-to-one correspondence between LiDAR points and image pixels based on the sensor calibration. These methods concatenate the features of both modalities in a deeper feature extraction stage. This point-level fusion scheme further exploits the high-dimensional feature complementarity of bi-modality features. The point-level feature fusion schemes proposed by EPNet [Huang *et al.*, 2020], PointAugmenting [Wang *et al.*, 2021], and PointPainting [Vora *et al.*, 2020] show excellent 3D object detection performance on multiple datasets, respectively. Recently, Transformer-based soft feature association methods [Li *et al.*, 2022; Bai *et al.*, 2022] show state-of-the-art 3D object detection results. The performance of these methods has a higher ceiling. DeepFusion [Li *et al.*, 2022] uses LiDAR data as queries for alignment of bi-modality features at the mid-level, which greatly reduces the interference of noisy features. In this paper, we aim to fully release the potential of Transformer-based architectures in multi-modality feature fusion. Instead of trusting the Transformer consistently, a confidence score is learned for each point concerning the different modality features it is associated to. The proposed method can effectively cope with the case of sparse geometric information or image texture degradation.

3 Bi-modality Feature Fusion Network with Both Soft and Hard Associations

In this part, we will present the proposed 3D object detection network architecture, in which the designed bi-modality feature fusion module will be described in detail. The whole network structure is divided into three parts to be described separately: 1) Initial feature extraction network for point clouds and images; 2) Object query initialization and initial object bounding boxes generation; 3) LiDAR-Camera features fusion module with both soft and hard components.

3.1 Image and LiDAR Feature Extraction Network

Like the previous works [Yoo *et al.*, 2020; Yin *et al.*, 2021a], the input data for the model comes from six cameras and a rotating mechanical LiDAR. The 2D backbone for extracting image features in our model uses ResNet50 [He *et al.*, 2016]. 3D backbone uses PointPillars [Lang *et al.*, 2019] or Voxelnet [Zhou and Tuzel, 2018]. The generated 3D voxel space features are compressed into BEV space as shown in Figure 2.

3.2 Object Querie Generation and Transformer Decoder

Given six image features $\{I_f^i | I_f \in \mathbb{R}^{H\times W\times C}\}, i = 1...6\}$ and one LiDAR BEV feature $L_f \in \mathbb{R}^{X\times Y\times C}$, our primary goal is to estimate a heat map that characterizes the 3D location of objects in space. $H \times W$ and $X \times Y$ represent the size of image features and BEV feature map, respectively, and C is the number of channels of the features. From the experimental experience of previous method [Zhou *et al.*, 2022; Yin *et al.*, 2021a], high recall of object position detection

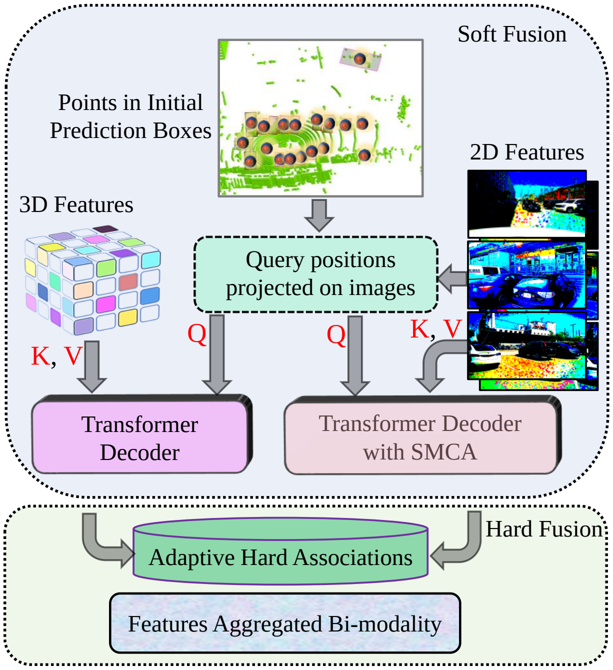

Figure 3: **We Propose A Bi-modality Feature Aggregation Scheme with Both Soft and Hard Components.** Initially the LiDAR points in the prediction boxes are projected onto the image plane through the camera parameters, which is what enable us to know which LiDAR points can be associated to pixels. The points that are considered valid are used as queries to perform soft feature association in both modalities. Each point softly associated to the two features is aggregated by the adaptive hard association module.

is already achievable with only LiDAR features. To further improve the recall of small object detection, we choose to use image feature-guided object heat map estimation, which is the soft attention feature association performed by camera features and LiDAR BEV features based on Transformer decoder [Bai *et al.*, 2022]. The design of Transformer decoder follows DETR [Carion *et al.*, 2020] and TransFusion [Bai *et al.*, 2022], we will present its detailed network structure in the supplementary material.

In the image-guided estimate object heat map module shown in Figure 2, LiDAR BEV features $L_f \in \mathbb{R}^{X\times Y\times C}$ are used as queries and we collapse the image features along the height axis $I_f^{\triangle} \in \mathbb{R}^{H\times C}$ as key values. The Spatially Modulated Cross Attention (SMCA) module proposed by TransFusion [Bai *et al.*, 2022] is used to implement feature interaction and construct bi-modality soft associations. $L_f \in \mathbb{R}^{X\times Y\times C}$ and $I_f^{\triangle} \in \mathbb{R}^{H\times C}$ are fed into the Transformer decoder with SMCA to obtain the updated features. The object heatmap generation can be represented by the following formula:

$$Heatmap = \frac{\Upsilon_{img}(I_f^{\triangle}) \oplus \Upsilon_{LiDAR}(L_f^{\bowtie})}{2}. \quad (1)$$

Υ_{img} and Υ_{LiDAR} in formula (1) represent the object heat map head for LiDAR BEV features and the object heat map head for image guidance, respectively, which is essentially a 2D convolution process. $L_f^{\bowtie}$ comes from reshaping LiDAR BEV features into $N \times C$ dimensional features. $\oplus$ represents the element-by-element summation. The generated heatmap $H \in \mathbb{R}^{X\times Y\times U}$ serves as our initial object query, where U is all the categories to be detected by the network. For object query initialization, the local maximum element in the heatmap H is considered as the initial position of the object.

As shown in Figure 2, the initial object query is input to the Transformer decoder as query positions. The original BEV feature is regarded as Key-Value. Taking advantage of the powerful attention mechanism in Transformer, the long-range dependence between objects in 3D space is modeled. Based on this, the feature representation F_{obj}^{init} of each initial object is updated and the soft association between features is learned. Then, we use the 3D detection head to regress the location information, size information, and category information of each object box. The formula is as follows:

$$\begin{aligned}\{\delta x^i, \delta y^i, log(l)^i, log(w)^i, log(h)^i, \sin\check{\theta}^i,\\ \cos\check{\theta}^i, {P_{obj}}^i | i = 0...s, P_{obj} \in [0,1]\} = \quad (2)\\ \Phi(TransD(Q = \ddot{H}, K = F_{obj}^{init}, V = F_{obj}^{init}),\end{aligned}$$

where x, y represent the centroids of the estimated objects. The length, width, and height of the 3D object box respectively are l, w, and h. We calculate the orientation of the front of the estimated object using the yaw angle $\check{\theta}^i$. $P_{obj} \in [0,1]$ represents the probability that the object is of each semantic class. s represents the number of estimated object positions. The local maximum element in the heat map is computed to generate the initial object queries $\ddot{H}$.

3.3 LiDAR-Camera Fusion Module with Both Soft and Hard Feature Associations

Given LiDAR BEV features $L_f \in \mathbb{R}^{X\times Y\times C}$, texture features $I_f \in \mathbb{R}^{H\times W\times C}$ of each camera, and initial object boxes information B_{inf}^{init}, how to better refine the coarse predictions B_{inf}^{init} of the first stage during the second stage is a challenge to be addressed. The bi-modality feature fusion on a point-wise level is one LiDAR point feature corresponding to one pixel point feature. This point-to-point feature fusion strategy [Huang *et al.*, 2020; Wang *et al.*, 2021] is not good at alleviating the disadvantage of LiDAR point cloud sparsity. It is because when only a few LiDAR point features are available at the estimated initial query location, such hard association fusion strategy only fuses a few pixel features as well. Existing Transformer-based cross-attention fusion methods [Li *et al.*, 2022; Bai *et al.*, 2022] provide good mitigation of the waste of high-resolution camera features. Although the Transformer-based soft feature fusion methods achieve what information should be obtained from the image, this schemes is unable to provide a good confidence score of its fused image features relative to the LiDAR BEV features.

Based on above insight, we propose a bi-modality fusion strategy with both soft and hard components. As shown in Figure 3, the object center points $Cen_p\{x, y\}$ obtained in the first stage are used as queries Q. On the one hand, the global LiDAR features of each center point Cen_p are further updated based on the Transformer decoder to model the geometric dependency between the objects in 3D space. On the other

Figure 4: **Objects that are difficult to detect in images or in point clouds.** The LiDAR point cloud is projected onto the image plane and shown in the lower left corner. In a 3D object detection network with bi-modality fusion, the feature representation should be dominated by whichever sensor signal is more effective. The attention mechanism of transformers enable the network to learn a better global representation. Combined with the transformers, our method gives the network the ability to adaptively recall better quality features. The proposed scheme achieves high quality 3D detection even when one sensor signal is poor.

hand, the Transformer-based decoder with SMCA [Bai *et al.*, 2022] is used to cross-attention image features. Since each Cen_p is soft-associated with a global 360-degree LiDAR features, the more reasonable method is that the soft-associated fused image features should also be global 360-degree all-views. The image features of the six camera views are collated as a complete texture feature library. Each center points query the global image features from the texture feature library. Then, we estimate a confidence score ($S^L_{conf} \in [0,1]$, $S^I_{conf} \in [0,1]$) for each center point's learned LiDAR features and image features separately, which score represents which LiDAR features and image features are more reliable. Finally the features $F_{S\&H}$ of each center point are adaptively aggregated by the confidence score.

In the adaptive hard fusion module of Figure 3, inspired by EPNet [Huang *et al.*, 2020], The designed network estimates the confidence scores of the two modal features separately for each point. The specific process is represented by the following formula:

$$S^L_{conf} = Sigmoid(\mathcal{F}(tanh(\mathcal{F}(L^{\bowtie}_f) + \mathcal{F}(I^{\Delta}_f)))), \quad (3)$$

where $Sigmoid$ means Sigmoid function. $\mathcal{F}$ means multi-layer perceptron (MLP), and $tanh$ is the tangent function. The confidence score S^I_{conf} of the image features is $1-S^L_{conf}$. Completing the second stage of updating the center point features, we refine the object detection information using the 3D detection head mentioned in Section 3.2. The details of the loss function of our 3D object detection network are presented in the supplementary material.

In Figure 4, in the pink frame, the profile and semantic information of the pedestrian can be clearly obtained from the image. In contrast, only one or two LiDAR points fall on the pedestrian with black clothes in the pink frame, and there are no LiDAR points at all on the lower half of his body. In this case, the previous hard fusion strategy only associates a very small number of pixel features causing the loss of more high-quality texture features. In contrast, the Transformer-based soft fusion strategy enables the association of rich image features, but this scheme is not perceived the importance of image features relative to LiDAR features. In the green box, the LiDAR points that fall more on the pedestrian provide better information about the geometric contour of the pedestrian. In this case, although the previous hard fusion strategy associates more texture features, these features are not effective enough for 3D detection. Our proposed approach intends to estimate the confidence score of the cross-modal. Such a method enables the network to determine which current modal features are more effective during loss backpropagation between predictions and ground truth, which then gives the network a robust learning capability. High quality camera texture features but sparse LiDAR points lead to loss of pedestrian geometric profile. Image quality is degraded due to strong nighttime light but LiDAR point cloud is better at drawing the pedestrian profile. These are two quite common autonomous driving scenarios. For these two cases, our model provides higher attention weights for the higher quality features. As shown in the upper right corner in Figure 4, our method shows robust performance in all cases.

4 Experiments

4.1 Settings

Datasets for Model Training and Evaluation

NuScenes [Caesar *et al.*, 2020] is a prestigious real-world dataset of autonomous driving scenes. nuScenes acquisition vehicles are equipped with one spinning LiDAR, five long range RADAR sensors and six cameras, which contribute significantly to the development of autonomous driving algorithms. A 32-beam LiDAR captures point cloud data at a frequency of 20 HZ. 6 surround-view cameras cover a 360-degree scene with no dead angle. nuScenes dataset [Caesar *et al.*, 2020] contains 700 training scenes, 150 validation scenes and 150 test scenes.

Evaluation Metric

The nuScenes dataset provides a variety of metrics for evaluating model performance. mean Average Precision (mAP): 2D Euclidean center distance error for 2D center points under BEV, different from IoU in KITTI [Geiger *et al.*, 2013]. nuScenes Detection Score (NDS): Weighted average of multiple evaluation metrics for the nuSceness dataset [Caesar *et al.*, 2020]. Average Translation Error (mATE): Average distance error of center point. Average Scale Error (mASE): Average scale error with object center point and orientation alignment. Average Orientation Error (mAOE): Average orientation error between predicted and ground truth.

Implementation Details

The network proposed in this paper is based on the popular general-purpose 3D object detection platform MMDetection3D [Contributors, 2020] on the PyTorch framework to build the model. We select two 3D backbones, VoxelNet

Table 1: **Comparison of our method with the best method on the popular nuScenes test set.** "Fusion" represents the feature fusion scheme of the methods, where "S&H" represents our proposed fusion scheme of both hard and soft components. "NO" represents the LiDAR-only methods. The table shows the evaluation results of the average detection precision for the ten categories in the nuScenes dataset. We also divide the ten categories into two groups: large objects and small objects, where "C.V.", "Ped." and "T.C." represent construction vehicles, pedestrians and traffic cones, respectively. "TransFusion (P)" represents the network designed by TransFusion based on the PointPillar feature extraction structure. We present better evaluation results for TransFusion (P) than in its paper. "FusionPainting (P)" represents the PointPillar baseline-based method of FusionPainting. The bolded font in the table indicates the optimal result for each part.

Method	Fusion			Relatively Large Objects					Relatively Small Objects				
		mAP	NDS	Car	Truck	C.V.	Bus	Trailer	Barrier	Motor.	Bike	Ped.	T.C.
PointPillar [Lang *et al.*, 2019]	NO	40.1	55.0	76.0	31.0	11.3	32.1	36.6	56.4	34.2	14.0	64.0	45.6
PointPainting [Vora *et al.*, 2020]	Hard	46.4	58.1	77.9	35.8	15.8	36.2	37.3	60.2	41.5	24.1	73.3	62.4
FusionPainting (P) [Xu *et al.*, 2021]	Hard	60.7	66.0	83.5	56.9	21.6	69.9	39.2	58.4	**66.4**	**54.1**	82.9	74.5
TransFusion (P) [Bai *et al.*, 2022]	Soft	59.6	65.4	**86.5**	**58.3**	23.4	**71.2**	38.7	60.0	66.0	44.0	82.8	65.1
Ours (PointPillar)	NO	55.0	62.8	84.7	55.2	21.3	67.4	37.0	60.3	57.4	30.6	79.3	56.6
Ours (PointPillar)	S&H	**62.1**	**66.7**	86.0	53.9	**32.0**	62.8	**56.2**	**65.0**	63.8	41.6	**83.8**	**77.2**
3DCVF [Yoo *et al.*, 2020]	Hard	52.7	62.3	83.0	45.0	15.9	48.8	49.6	65.9	51.2	30.4	74.2	62.9
MVP [Yin *et al.*, 2021b]	Hard	66.4	70.5	86.8	58.5	26.1	67.4	57.3	**74.8**	70.0	49.3	**89.1**	85.0
FusionPainting [Xu *et al.*, 2021]	Hard	66.5	70.6	87.0	**62.9**	25.3	**70.6**	45.0	67.2	**74.6**	**64.4**	88.4	79.5
AutoAlign [Chen *et al.*, 2022]	Hard	65.8	**70.9**	85.9	55.3	29.6	67.7	55.6	-	71.5	51.5	86.4	-
LIFT [Zeng *et al.*, 2022]	Soft	65.1	70.2	**87.7**	55.1	29.4	62.4	59.3	69.3	70.8	47.7	86.1	83.2
Ours (VoxelNet)	S&H	**67.6**	**71.0**	**87.7**	58.1	**32.9**	68.0	**61.2**	74.0	71.3	50.0	88.0	**85.1**
TCT [Yuan *et al.*, 2022]	NO	50.5	-	83.2	51.5	15.6	63.7	33.0	53.8	54.0	**53.8**	74.9	52.5
CenterPoint [Yin *et al.*, 2021a]	NO	60.3	67.3	85.2	53.5	20.0	63.6	56.0	71.1	59.5	30.7	84.6	78.4
multi-task [Fazlali *et al.*, 2022]	NO	60.9	67.3	84.6	50.0	23.4	63.2	55.3	68.2	65.1	38.9	83.7	76.8
Afdetv2 [Hu *et al.*, 2022]	NO	62.4	68.5	86.3	54.2	26.7	62.5	58.9	71.0	63.8	34.3	85.8	80.1
S2M2-SSD [Zheng *et al.*, 2022]	NO	62.9	69.3	86.3	56.0	26.2	65.4	59.8	**75.1**	61.6	36.4	84.6	77.7
Ours (VoxelNet)	NO	**66.1**	**70.1**	**86.8**	**57.4**	**31.9**	**68.0**	**61.7**	74.4	**68.4**	43.1	**86.9**	**82.6**

[Zhou and Tuzel, 2018] and PointPillar [Lang *et al.*, 2019], as LiDAR feature extractors. A pre-trained ResNet50 [He *et al.*, 2016] is used as the 2D backbone to extract the image features. As with TransFusion [Bai *et al.*, 2022] and PointAugmenting [Wang *et al.*, 2021], we set the resolution of the images to 448×800 to reduce the training and inference time consumption. The main training steps consist of three: 1) Firstly, the 3D backbone, the first Transformer decoder and the 3D detection head are pre-trained with only LiDAR signals as input. About 20 epochs enable full convergence. In the first step we use the same data enhancement strategy as CenterPoint [Yin *et al.*, 2021a]; 2) Following the TransFusion training plan, we continue training for about 6 epochs in the second step without the SECOND [Yan *et al.*, 2018] data enhancement strategy; 3) In the third step, we train the image-guided estimation of the object heat map module and the bi-modality soft & hard fusion module based on 3D box center point guidance. We used two NVIDIA Quadro RTX 8000 with a batch size of 12 to train the neural network.

4.2 Experimental Results and Analysis

Comparison with The SOTA Baselines

All experimental results in this paper are submitted to the official nuScenes evaluation site for model performance evaluation. In Table I, we compare the state-of-the-art 3D object detection algorithms on the nuScenes ranking [Caesar *et al.*, 2020]. We also split the 10 categories of the comparison into two broad categories. Baseline PointPillar [Lang *et al.*, 2019] is mainly used in the upper part of Table 1. PointPainting [Vora *et al.*, 2020] performs hard bi-modality feature aggregation based on PointPillar. Inspired by Transformer's multi-headed attention mechanism, TransFusion [Bai *et al.*, 2022] performs soft bimodality feature aggregation based on PointPillar. Both fusion strategies have more significant performance improvements over the baseline PointPillar. The proposed bi-modality fusion strategy with both soft and hard fusion in this paper completely outperforms previous soft and hard fusion methods [Xu *et al.*, 2021; Vora *et al.*, 2020; Bai *et al.*, 2022] in both mAP and NDS metrics. In particular, significant accuracy improvements are achieved for uncommon classes (e.g., construction vehicle, trailer, and barrier) and small objects (pedestrian and traffic cone). Experimental results using PointPillar as a baseline demonstrate that the proposed combined soft and hard multi-modality fusion method outperforms either hard or soft fusion only.

In the middle part of Table 1, the main focus is on comparing the methods using non-PointPillar baselines. Our proposed method based on 3D backbone VoxelNet [Zhou and Tuzel, 2018] achieves a very outstanding overall performance. Comparing with state-of-the-art 3D object detection methods [Yoo *et al.*, 2020; Yin *et al.*, 2021b; Chen *et al.*, 2022; Xu *et al.*, 2021; Zeng *et al.*, 2022] with 2D-3D feature fusion, the method in this paper achieves the best performance in both mAP and NDS metrics. We attribute the proposed soft and hard bi-modality fusion strategy to play a

Table 2: **Comparison of different Backbone and different feature fusion schemes.** For PointPillars, the scheme of TransFusion is used for soft fusion and the scheme of PointPainting is used for hard fusion. For VoxelNet, the TransFusion scheme is used for soft fusion, and the direct feature concat method and PointAugmenting scheme are used for hard fusion. "Improvement" represents the increase in gain of our method compared to the baseline.

Method	Backbone	mAP	NDS
PointPillar	PointPillars	40.1	55.0
CenterPoint	PointPillars	50.3	60.2
TransFusion (LiDAR-Only)	PointPillars	54.5	62.7
Ours (LiDAR-Only)	PointPillars	55.0	62.8
Soft Fusion (TransFusion)	PointPillars	58.3	64.5
Hard Fusion (PointPainting)	PointPillars	46.4	58.1
Ours Fusion	PointPillars	**60.7**	**66.0**
Improvement$\uparrow$	None	**20.6**	**11.0**
CenterPoint	VoxelNet	59.6	66.8
Ours (LiDAR-Only)	VoxelNet	66.1	70.1
Soft Fusion (TransFusion)	VoxelNet	65.6	69.7
Hard Fusion (Point-wise concat)	VoxelNet	63.3	67.6
Hard Fusion (PointAugmenting)	VoxelNet	64.2	68.7
Ours Fusion	VoxelNet	**67.6**	**71.0**
Improvement$\uparrow$	None	**7.9**	**4.1**

Table 3: **Comparison of the detection errors of the different methods.** Our network has the smallest Average Translation Error (mATE) and Average Orientation Error (mAOE) compared to the SOAT method.

Method	Fusion	mATE$\downarrow$	mASE$\downarrow$	mAOE$\downarrow$
CenterPoint [Yin *et al.*, 2021a]	NO	0.262	0.239	0.361
Ours (PointPillars)	NO	0.332	0.277	**0.352**
Ours (PointPillars)	S&H	**0.296**	**0.250**	0.432
3DCVF [Yoo *et al.*, 2020]	Hard	0.300	0.245	0.458
TransFusion [Bai *et al.*, 2022]	Soft	0.259	**0.243**	0.359
Ours (VoxelNet)	NO	0.259	0.246	0.386
Ours (VoxelNet)	S&H	**0.258**	0.256	**0.356**

significant role. The bottleneck in the accuracy improvement of current multi-modality fusion methods is the poor detection accuracy of small objects and the non-robust perception performance for poor scenes. The strategy of combining both soft and hard fusion proposed in this paper is considered as a promising thought, especially for extreme environments. The performance of our LiDAR-only signal methods is also reported in the lower part of the Table 1. The LiDAR signal-only method in this paper achieves the best performance in almost all metrics.

Comparison of Different Feature Fusion Schemes

In Table 2, we report our results based on different backbone networks [Lang *et al.*, 2019; Zhou and Tuzel, 2018]. All the predict results in the table are submitted on the official nuScenes evaluation web site. Compared to the baseline PointPillar, our fusion strategy improves the mAP and NDS metrics by 20.6 and 11.0, respectively. Compared to existing 3D object detection algorithms with soft feature fusion only or hard feature fusion only, our proposed method demonstrates promising evaluation results. For the mAP metric, our algorithm outperforms PointPainting [Vora *et al.*, 2020], a hard fusion strategy, by about 30%. Also, the proposed algorithm outperforms TransFusion [Bai *et al.*, 2022], a soft fusion strategy, by about 4%. For the VoxelNet backbone , compared to CenterPoint [Yin *et al.*, 2021a] using VoxelNet as the feature extractor, our method improves the mAP and NDS metrics by 7.9 and 4.1, respectively. It is noticed from Table II that both soft feature fusion-only and hard feature fusion-only strategies have shown great gains for 3D object detection tasks. However, the performance improvement of 3D object detection networks with soft-feature fusion only or hard-feature fusion is already facing increasingly difficult breakthroughs. Previous methods [Jiang *et al.*, 2022; Huang *et al.*, 2020; Zeng *et al.*, 2022; Yoo *et al.*, 2020] have also corroborated the effectiveness of both strategies one by one. Based on this insight, it is reasonable solution how to better combine the advantages of point-by-point hard feature fusion and soft association fusion based on Transformer's multi-headed attention mechanism. To design a fusion strategy that takes into account the advantages of both will create a breakthrough for robots to perceive the 3D real world more accurately. In this paper , we explore such a scheme and demonstrate its effectiveness through various experiments on the nuScenes dataset [Caesar *et al.*, 2020].

Error Analysis of Model Prediction Results

We report the results of the evaluation of our method on mATE, mASE and mAOE metrics in Table 3. Likewise, this three evaluation metrics measure the robustness of the model for perception of multiple scenes. Our PointPillar-based fusion method shows significant improvement in the mATE and mASE metrics. Compared to other fusion methods [Bai *et al.*, 2022; Yoo *et al.*, 2020], our VoxelNet-based approach demonstrates low errors on mATE and mAOE. The effectiveness of the bi-modality fusion strategy with both hard and soft is confirmed by these metrics.

5 Conclusion

In this paper, we reveal the key challenges facing LiDAR feature and camera LiDAR feature fusion. To bridge the disadvantages of previous methods for both hard and soft feature association, we propose a bi-modality feature fusion strategy with both soft and hard components. Comparison experiments demonstrate the effectiveness of our idea. The designed network alleviates the performance degradation of the network caused by poor features of individual sensor signals and improves the detection precision of small objects. The model proposed in this paper shows competitive results on the nuScenes dataset. In particular, optimal prediction results are achieved for the estimation of challenging categories such as trailer, construction vehicle and traffic cone. We believe that such bi-modality fusion strategy of both hard and soft will be beneficial to other fields as well.

References

[Bai *et al.*, 2022] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan 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 (CVPR)*, pages 1090–1099, June 2022.

[Caesar *et al.*, 2020] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[Carion *et al.*, 2020] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[Chen *et al.*, 2017] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017.

[Chen *et al.*, 2022] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinghong Jiang, Feng Zhao, Bolei Zhou, and Hang Zhao. Autoalign: Pixel-instance feature aggregation for multi-modal 3d object detection. *arXiv preprint arXiv:2201.06493*, 2022.

[Contributors, 2020] MMDetection3D Contributors. Mmdetection3d: Openmmlab next-generation platform for general 3d object detection, 2020.

[Fazlali *et al.*, 2022] Hamidreza Fazlali, Yixuan Xu, Yuan Ren, and Bingbing Liu. A versatile multi-view framework for lidar-based 3d object detection with guidance from panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17192–17201, 2022.

[Geiger *et al.*, 2013] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Hu *et al.*, 2022] Yihan Hu, Zhuangzhuang Ding, Runzhou Ge, Wenxin Shao, Li Huang, Kun Li, and Qiang Liu. Afdetv2: Rethinking the necessity of the second stage for object detection from point clouds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 969–979, 2022.

[Huang *et al.*, 2020] Tengteng Huang, Zhe Liu, Xiwu Chen, and Xiang Bai. Epnet: Enhancing point features with image semantics for 3d object detection. In *European Conference on Computer Vision*, pages 35–52. Springer, 2020.

[Jiang *et al.*, 2022] Chaokang Jiang, Guangming Wang, Jinxing Wu, Yanzi Miao, and Hesheng Wang. Ffpa-net: Efficient feature fusion with projection awareness for 3d object detection. *arXiv preprint arXiv:2209.07419*, 2022.

[Ku *et al.*, 2018] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven L Waslander. Joint 3d proposal generation and object detection from view aggregation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–8. IEEE, 2018.

[Lang *et al.*, 2019] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019.

[Li *et al.*, 2022] Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V Le, et al. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17182–17191, 2022.

[Qi *et al.*, 2017] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas 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*, pages 652–660, 2017.

[Qi *et al.*, 2018] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 918–927, 2018.

[Roy *et al.*, 2022] Debashri Roy, Yuanyuan Li, Tong Jian, Peng Tian, Kaushik Roy Chowdhury, and Stratis Ioannidis. Multi-modality sensing and data fusion for multi-vehicle detection. *IEEE Transactions on Multimedia*, 2022.

[Sagar, 2022] Abhinav Sagar. Aa3dnet: attention augmented real time 3d object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 628–635, 2022.

[Shin *et al.*, 2019] Kiwoo Shin, Youngwook Paul Kwon, and Masayoshi Tomizuka. Roarnet: A robust 3d object detection based on region approximation refinement. In *2019 IEEE intelligent vehicles symposium (IV)*, pages 2510–2515. IEEE, 2019.

[Sun *et al.*, 2020] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020.

[Vora *et al.*, 2020] Sourabh Vora, Alex H. Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[Wang and Jia, 2019] Zhixin Wang and Kui 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)*, pages 1742–1749, 2019.

[Wang *et al.*, 2021] Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang. Pointaugmenting: Cross-modal augmentation for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11794–11803, 2021.

[Xu *et al.*, 2021] Shaoqing Xu, Dingfu Zhou, Jin Fang, Junbo Yin, Zhou Bin, and Liangjun Zhang. Fusionpainting: Multimodal fusion with adaptive attention for 3d object detection. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 3047–3054. IEEE, 2021.

[Yan *et al.*, 2018] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018.

[Yin *et al.*, 2021a] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021.

[Yin *et al.*, 2021b] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Multimodal virtual point 3d detection. *Advances in Neural Information Processing Systems*, 34:16494–16507, 2021.

[Yoo *et al.*, 2020] Jin Hyeok Yoo, Yecheol Kim, Jisong Kim, and Jun Won Choi. 3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection. In *European Conference on Computer Vision*, pages 720–736. Springer, 2020.

[Yuan *et al.*, 2022] Zhenxun Yuan, Xiao Song, Lei Bai, Zhe Wang, and Wanli Ouyang. Temporal-channel transformer for 3d lidar-based video object detection for autonomous driving. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4):2068–2078, 2022.

[Zeng *et al.*, 2022] Yihan Zeng, Da Zhang, Chunwei Wang, Zhenwei Miao, Ting Liu, Xin Zhan, Dayang Hao, and Chao Ma. Lift: Learning 4d lidar image fusion transformer for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17172–17181, June 2022.

[Zheng *et al.*, 2022] Wu Zheng, Mingxuan Hong, Li Jiang, and Chi-Wing Fu. Boosting 3d object detection by simulating multimodality on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13638–13647, 2022.

[Zhou and Tuzel, 2018] Yin Zhou and Oncel 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*, pages 4490–4499, 2018.

[Zhou *et al.*, 2022] Zixiang Zhou, Xiangchen Zhao, Yu Wang, Panqu Wang, and Hassan Foroosh. Centerformer: Center-based transformer for 3d object detection. In *European Conference on Computer Vision*, pages 496–513. Springer, 2022.