IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 25 May 2023, accepted 21 July 2023, date of publication 1 August 2023, date of current version 10 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3300708

RESEARCH ARTICLE

Monocular 3D Object Detection With Motion Feature Distillation

HENAN HU[1,2], MUYU LI[3], MING ZHU[1], WEN GAO[4], PEIYU LIU[5], AND KWOK-LEUNG CHAN[6]

[1]Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]Centre for Intelligent Multidimensional Data Analysis Ltd., Hong Kong, China
[4]BYD Auto Industry Company Ltd., Shenzhen 518119, China
[5]Shenyang Aircraft Design & Research Institute, Shenyang, Liaoning 110036, China
[6]Department of Electrical Engineering, City University of Hong Kong, Hong Kong, China

Corresponding author: Kwok-Leung Chan (itklchan@cityu.edu.hk)

This work was supported in part by the Hong Kong Innovation and Technology Commission [InnoHK Project Centre for Intelligent Multimedia Data Analysis Limited (CIMDA)], City University of Hong Kong Strategic Research Grant 7005855, and in part by the Jilin Scientific and Technological Development Program under Grant 20220201146GX.

ABSTRACT In the context of autonomous driving, environmental perception within a 360-degree field of view is extremely important. This can be achieved via the detection of three-dimensional (3D) objects in the surrounding scene with the inputs acquired by sensors such as LiDAR or RGB camera. The 3D perception generated is commonly represented as the bird's-eye-view (BEV) of the sensor. RGB camera has the advantages of low-cost and long-range acquisition. As the RGB images are two-dimensional (2D), the BEV generated from 2D images suffers from low accuracy due to limitations such as lack of temporal correlation. To address the problems, we propose a monocular 3D object detection method based on long short-term feature fusion and motion feature distillation. Long short-term temporal features are extracted with different feature map resolutions. The motion features and depth information are combined and encoded using an encoder based on the Transformer cross-correlation module, and further integrated into the BEV space of fused long short-term temporal features. Subsequently, a decoder with motion feature distillation is used to localize objects in 3D space. By combining BEV feature representations of different time steps, and supplemented with embedded motion features and depth information, our proposed method significantly improves the accuracy of monocular 3D object detection as demonstrated from experimental results obtained on nuScenes dataset. Our proposed method outperforms state-of-the-art methods, in particular the previous best art by 6.7% on mAP, and 8.3% on mATE.

INDEX TERMS 3D object detection, bird's-eye-view (BEV), monocular depth estimation, motion feature, knowledge distillation, autonomous driving.

I. INTRODUCTION

In applications such as autonomous driving [1], [2], [3], robot navigation, and augmented reality, any environmental changes within the 360-degree field of view can directly affect the safety of the vehicle and the correctness of self-driving decisions. Therefore, accurate three-dimensional (3D) perception of the surrounding scene is important and crucial. In order to achieve accurate 3D perception and objects localization, previous methods use sensors such as LiDAR or panoramic cameras to provide accurate distance measurements and algorithms adapted to the corresponding sensors. However, the increasing demand of additional sensors results in higher deployment costs and algorithm complexity, greatly affecting the computational cost and generalization ability of downstream tasks such as 3D object detection, object tracking, and 3D instance segmentation. To solve these problems, low-cost and easy-to-deploy monocular 3D object detection has become a feasible solution and research direction for panoramic scene perception.

Monocular 3D object detection, with only inputs of RGB two-dimensional (2D) images, lacks additional sensors that can provide accurate depth information. The ambiguity of depth values makes the monocular 3D object detection an

The associate editor coordinating the review of this manuscript and approving it for publication was Szidonia Lefkovits.

ill-posed problem. Therefore, using 2D images as the only information source suffers from low accuracy in estimating the distance of objects and predicting the 3D bounding boxes. In recent years, some works [4], [5], [6], [7], [8], [9] have used short-term neighboring image information, e.g., temporal window, for detecting and locating 3D objects from 2D images of the scene. It is expected that utilizing a larger temporal window can bring greater disparity in dynamic circumstances. Accuracy of estimating object distance, and subsequently the 3D object detection, will be improved. However, due to high computational cost and concerns in deploying the model, these works only exploit 2-3 frame temporal window in the neighborhood to provide temporal information support. The difficulties and the demand for higher accuracy are still faced by the current research in monocular 3D object detection.

Existing methods have tried different ways to aggregate temporal features for locating objects in 3D space. Generally, these methods adopt spatio-temporal stereo matching. The algorithm is to process temporal information in a virtual 3D space and simultaneously consider image features corresponding to spatial hypothetical positions generated from multiple time points. To quantify the quality of these methods in stereo depth estimation, Park et al. [10] proposed 3D localization potential which is defined as the amplitude of the change in the projected length of a source view caused by a change in the depth of a reference view. A larger 3D localization potential causes the corresponding pixel in the reference view to be projected further in the source view, providing source view features with greater differentiation for depth estimation in the reference view. Therefore, depth estimation values that are more strongly associated with source view features can suppress incorrect depth estimation and result in more accurate 3D spatial localization of objects. However, if temporal window is small, 3D localization potential will be low. This will lead to lower accuracy of 3D object localization. Another factor is the resolution of the input image feature map. Lower feature map resolution can also limit the accuracy of depth estimation and affect the performance of 3D object detection.

Based on the above discussion, we first propose a long-short fusion idea for monocular 3D object detection based on an expanded temporal window. Different parameters are selected to balance the feature map resolution and temporal window size via two designs: (i) improving the feature map resolution of the input images in the temporal neighborhood of the source view; and (ii) reducing the feature map resolution of the images further away in the temporal domain. As a result, our proposed method, with the use of a larger temporal window, can exploit richer temporal features from more reference views. With larger 3D localization potential, accuracy of 3D object detection is improved.

When the motion pattern of the observation point tends to be stable, a larger temporal window can provide reliable temporal features. However, in real-world scenarios, motion pattern of the observation point may be randomly disturbed due to changes of the surrounding environment, thereby affecting the algorithm's ability of modeling temporal information. Therefore, we introduce motion estimation as an auxiliary process which is utilized to extract motion features for 3D object detection. However, motion estimation from 2D images is not a trivial task. Errors can cause feature blurring and instability of the object detection model, leading to inaccurate 3D object localization. To solve the above problems, we propose our monocular 3D object detection method by combining the long-short fusion idea with motion feature distillation. Unlike other methods that use binocular vision or LiDAR input data for knowledge distillation, our distillation process uses a unified teacher-student structure which is trained with the motion state calculated from the target depth ground truth as the input of the teacher model. A unified model helps aligning feature space and response to avoid potential errors between the teacher and student models. Specifically, the motion feature distillation process performs knowledge distillation on the intermediate features and responses of the model, using Transformer to capture the global motion feature correlation in the temporal domain. The self-attention mechanism of the 3D motion feature perception interacts and fuses the positional encoding based on the ground truth motion value with the semantic features of the input image. The teacher model can provide additional regularization for the student model to reduce the impact of motion estimation errors on 3D object localization. During the model inference process, the student model can completely discard the positional encoding to avoid dependence on ground truth motion information.

Our contributions are summarized as follows:

- We propose the long-short fusion idea for monocular 3D object detection method. It balances the demands for larger temporal window and constraint on model complexity. An expanded temporal window provides higher feature map resolution of the temporal neighborhood. Reducing the feature map resolution of the images further away in the temporal domain helps to avoid the large increment of the number of model parameters. Therefore, performance of our proposed monocular 3D object detection is enhanced with the utilization of rich temporal features, while the model is still efficient by avoiding significant increase of memory consumption.
- Motion features can be used for monocular depth estimation. We propose the motion feature distillation method to tackle the problem of temporal feature pollution caused by random motion of the observation point. The enhanced motion features help to improve the stability of 3D object localization and the accuracy of depth estimation. To the best of our knowledge, our proposed monocular 3D object detection method is the first to utilize knowledge distillation to provide improved motion features for the depth estimation process.
- We propose the monocular 3D object detection with the novel designs of long-short fusion and motion feature

distillation. Experiments has been performed on the nuScenes dataset. We evaluate quantitatively as well as visualize the quality of 3D object detection in BEV space. We also study the robustness of the proposed model. In comparative analysis, we demonstrate that our proposed method surpasses state-of-the-art methods by 6.7% or more on mAP, 6.9% or more on NDS, and also on other evaluation metrics.

Our paper is structured as follows. The related studies are reviewed in Section II. We focus on various monocular 3D object detection models. In addition, previous works about knowledge distillation are introduced. Section III describes our proposed framework for 3D object detection. We evaluate our framework and compare its performance with state-of-the art methods. Section IV presents the experimental results and comparative analysis. Finally, in Section V, we draw the conclusion and outline future research directions.

II. RELATED WORK

RGB camera-based 3D object detection methods [11], [12], [13] have been proposed for tasks such as robot navigation, autonomous driving, and path planning. These methods, relying solely on information within the front view, are insufficient to provide a complete 3D perception of the environment. Therefore, utilizing 360-degree panoramic images that can provide comprehensive environmental information has gradually become a focus of attention in the field of 3D perception. Typically, the 360-degree panoramic information is composed of multiple non-overlapping RGB images. This kind of implementation is lower in cost than other sensors and has broad application prospects and high value.

A. MULTI-VIEW MONOCULAR 3D OBJECT DETECTION

In recent years, some works [4], [14], [15] have started to focus on the research of multi-view monocular 3D object detection tasks. Most monocular 3D object detection algorithms operate in a 3D space centered on the observation point. They map the multi-view RGB image features covering a 360-degree range through deep network to 3D space, and understand the surrounding environment in this space. To address the uncertainty of depth information for monocular sensors, some methods introduce neighboring frames to provide temporal information to the model, and improve depth estimation accuracy by using the disparity produced by the temporal motion of the targets.

Monocular 3D object detection methods assisted by multi-frame information focus on different ways of extracting temporal information. However, these works generally include the following core parts: 1) candidate regions selection for 3D localization; 2) feature extraction for candidate regions; 3) selection of feature map resolution for sampling; 4) multi-frame feature fusion method; 5) selection of the number of neighboring frames in the temporal domain; 6) processing of temporal features when fusing candidate regions; 7) training methods for specific tasks.

Early work in the field of multi-view monocular 3D object detection began with MVSNet [5], which selects each spatial point in the reference view as a candidate region. For these spatial points, this method projects them onto each view to obtain corresponding image features using 1/4 resolution and bilinear sampling, and completes feature fusion. Then, 3D convolution is used to extract features for candidate positions in the scanning plane. The object detection process uses fused temporal features to predict the probability of the existence of targets in each candidate region. The probability distribution of each pixel on depth obtained by temporal feature prediction is used to weight and generate a single depth prediction for each pixel, which is then supervised by L_1 loss. MVSNet is able to reconstruct 3D scene from single image with different conditions (input resolution, lightning condition, and viewpoint). However, it is computationally expensive and can be sensitive to noise and complex geometry. The candidate positions in MaGNet [6] predict the spatial points within the Gaussian confidence interval of the predicted depth, which is similar to MVSNet, except that MaGNet uses dot product instead of variance to aggregate temporal features and uses 2D CNN to process temporal features of candidate positions. MaGNet iteratively updates the Gaussian distribution of depth using these processed features and aligns it with the ground truth depth using L_2 supervision.

Another type of work is the monocular 3D object detection method based on LSS [7], which projects the features of panoramic RGB images onto a BEV grid through depth information and camera parameters. Subsequently, a method based on 2D object detection is used to extract the objects to be detected in the BEV. The candidate positions of BEVDet4D [4] are BEV grid units, and each unit samples features by pooling spatial points located within each unit. BEVDet4D aligns the previous BEV feature map with the current one through the trend of the observation point's motion and stacks them together. In this way, each grid unit can receive image features from multiple time steps. Subsequent methods, such as BEVDepth [8] and BEVStereo [9], optimize the depth estimation and 3D localization accuracy of this method by introducing multiple timesteps image information and MVSNet's spatio-temporal depth estimation module. LSS-based methods fuse multi-frame features at low-resolution, but the number of timesteps used for temporal fusion are limited. Thus, influence the performance of depth inference.

With the rapid development of Transformer in the field of computer vision, some methods have attempted to introduce Transformer's self-attention module to improve the model's ability to model global correlations. BEVFormer [14] uses a query module with self-attention mechanism to aggregate image features at different time steps on the BEV grid and uses a Transformer decoder in the subsequent stage to obtain the final 3D localization result from the aggregated query feature vector. BEVFormer can store historical query feature vectors during the inference process to continuously fuse temporal features, but fusing more than four timesteps does

not improve the model's inference ability. Unlike other works that directly aggregate image features onto the BEV grid, PolarFormer [16] generates an intermediate Polar BEV representation based on the polar coordinate system to store the fused image features. This polar coordinate-based representation is closer to the imaging principle of the camera and is more direct for representing objects in the far distance of the image. PETRv2 [17] is different from the previous works in that it does not aggregate the image features corresponding to the candidate regions through projection or attention mechanism, but aggregates the temporal features by computing the unconstrained cross-attention between the previous and current image features. This work projects the image features outward and uses 3D position encoding to represent these features, and encourages the candidate location of the query to attend to the image features at any time step related to its space. Query-based methods cost more computational resources when using self- or cross-attention mechanism, thus limiting the potential of utilizing more temporal information.

Inspired by the fusion manner of SOLOFusion [10], we integrate motion analysis into the long and short-term temporal aggregation. With the motion pattern of each objects, our proposed model is able to enhance the temporal feature of each object, and also communicate different motion features from objects with different moving trend. Furthermore, considering the noise and the random movement of the viewpoint and inspired by knowledge distillation, we propose the motion feature distillation to address the temporal pollution problem and boost the temporal fusion performance.

B. KNOWLEDGE DISTILLATION

Knowledge Distillation (KD) was originally developed as a technique for model compression tasks [18]. It works by building a teacher network and a student network, allowing the student network to learn from both the ground truth and soft labels generated by the teacher network. The work [18] shows that the student network can also be guided by the intermediate layer features of the teacher network. Since then, many methods have successfully applied knowledge distillation techniques in various tasks, such as image classification [19], semantic segmentation [20], and depth estimation [21].

In the field of 2D object detection, the work [19] first used knowledge distillation by extracting features from the intermediate layers, detection regression heads, and object classification heads of the network. To solve the problem of foreground-background imbalance, some works simulate the detection regions of the student network from candidate regions or extract fine-grained features from foreground object areas for knowledge distillation. In the task of 3D object detection, LiGA-Stereo [22] guides the learning of a visual-based monocular 3D object detection student network using geometric features obtained from a LiDAR-based 3D object detector to alleviate the impact of depth estimation errors. Recently, MonoDistill [23] designed a teacher network based on LiDAR signals to train a student monocular 3D object detector with spatial clues. Yang et al. [24] proposed an improved KD pipeline which differs from previous methods by using knowledge distillation techniques on a series of intermediate features and network responses to transfer depth information from a teacher network trained with ground truth depth information to a student network. This method does not require additional multi-view stereo or LiDAR data as input, making it applicable to a wider range of scenarios.

This paper proposes a long short-term fusion BEV-based monocular 3D object detection method with an expanded temporal window to improve the performance of monocular 3D object detection. Unlike the aforementioned works, this paper also proposes a motion feature distillation method that uses the relative motion trend between the target in the scene and the observation point to compensate for the temporal feature pollution caused by irregular motion patterns and improves the spatial localization accuracy and stability of the BEV-based monocular 3D object detection model.

III. METHODOLOGY

Our proposed method attempts to enlarge the temporal window and extract image features from varying feature map resolution at different time steps. Meanwhile, to alleviate the temporal feature disturbance caused by the randomness of the observer's self-motion, a network training technique based on motion feature distillation is proposed to improve the stability of the network model regarding temporal features and the accuracy of BEV 3D localization.

A. LONG-SHORT SPATIO-TEMPORAL FUSION

The core of the proposed method for monocular 3D object detection based on long-short spatio-temporal fusion is to balance the resolution of feature maps and the performance impact of temporal feature fusion with an expanded temporal window for 3D object localization in a 360-degree environment. The overall method, as shown in Fig. 1, consists of two parts: 1) long-term low-resolution feature fusion based on LSS; 2) short-term high-resolution feature fusion based on MVSNet.

1) LONG-TERM LOW-RESOLUTION FEATURE MAP FUSION

To introduce a larger temporal window while reducing network computational burden, the image feature map resolution corresponding to time steps with larger time intervals will be compressed to 1/16 size. These compressed feature maps are then used to generate dense point cloud information together with the corresponding depth prediction results, which are then voxelized into BEV features. By introducing more time frames in the temporal domain and increasing the time interval with the source view, the potential degradation of 3D localization caused by low image resolution can be addressed. Specifically, 16 historical frames starting from the source view of the current frame are aligned to the current time node. All time nodes, together with the feature map of the

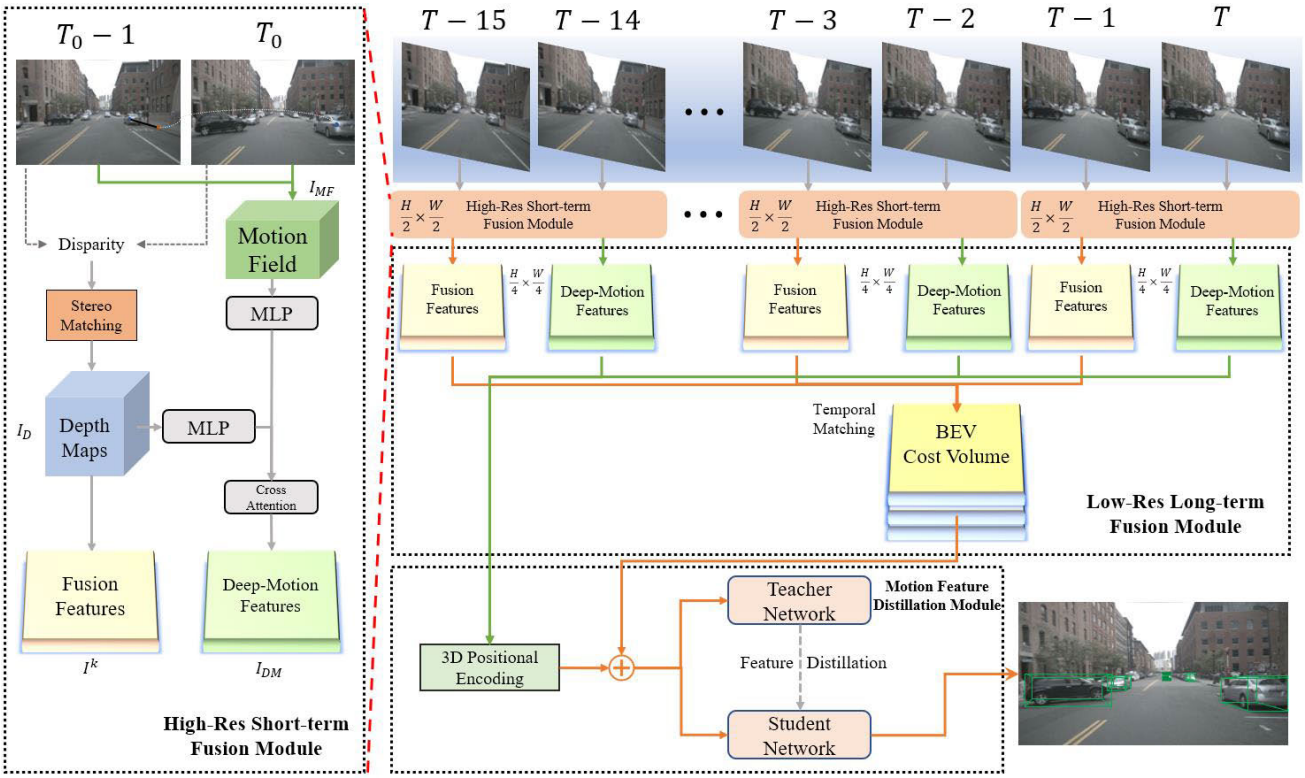

FIGURE 1. Overview of long short-term monocular 3D object detection.

current frame, are stacked, and then foreground object 3D localization is performed on this stacked feature vector.

2) SHORT-TERM HIGH-RESOLUTION FEATURE MAP FUSION

For the information of adjacent frames in the temporal domain, this method uses short-term high-resolution feature maps for compensation. On top of the above-mentioned long-range temporal matching fusion network, a depth matching model based on the MVSNet network structure is added for adjacent frames. For the temporal feature map of the current frame, this method selects a high-resolution (1/4 of original image resolution) top view feature to generate and performs stereo matching with the image feature of the previous frame. In the subsequent localization process, the monocular depth estimation module and the 3D spatial localization module interact with each other through cross-attention mechanism.

The MVSNet model estimates the depth value of a candidate region through 3D convolution. Estimating depth and 3D localization in the entire virtual 3D space requires a large amount of computational resources. For depth estimation in MVSNet, there is no prior knowledge, and a 128-dimensional discrete vector is required to cover the entire depth interval. This method is inspired by the balance between exploitation and exploration commonly seen in reinforcement learning, and the advantages of depth estimation and 3D localization to complement each other. Depth estimation provides prior knowledge to reduce the dimensionality of the feature space and lighten the computational cost of the model. If monocular depth estimation can be used as prior knowledge, the Top-k possible depth values of the candidate region are roughly estimated first, thereby reducing the dimensionality of the depth vector to k and reducing memory usage. This method first selects k monocular depth estimation values with higher confidence, reduces the weight of depth confidence in adjacent depth intervals, and forces the model to lower the probability of depth values near the Gaussian distribution of higher confidence depth values. This process is repeated for each pixel to produce a set of k candidate depths, and the subsequent MVSNet 3D localization process is performed on these points. This method can cover multi-modal depth distributions and reduce computational costs. Short-term high-resolution 3D localization module and long-term low-resolution 3D localization module complement each other and optimize the final 3D object detection results.

B. MULTI-FRAME IMAGE MOTION FEATURE DISTILLATION

When incorporating multiple frames of image information, if the motion state of the observation point is disturbed, the position of the target in the sampled multiple frames of images will undergo random changes, which affects the depth estimation of the target by the above network. At the same time, during the normal motion process of the observation point, there is no regularity in the changes of direction and speed. Directly learning from multiple frames of images is

difficult to accurately capture the temporal changes in features. To solve the problem of difficult model training due to the motion of the observation point, this method introduces knowledge distillation technology, and uses the motion state variation obtained from 3D annotation information of the foreground target in multiple frames of images to train a teacher network. The motion features are then distilled and transferred to the student network. The student network has the ability to compensate for motion features during inference, thereby improving the accuracy and stability of the network in determining the depth of the target.

The tokens for teacher and student model are designed as the fused intermediate feature maps with motion, depth and image information. The overall process of motion feature distillation is shown in Fig. 2. Firstly, the feature vectors $\{I^k\}_{k\in(1,\ldots,n)}$ and $\{T^k\}_{k\in(1,\ldots,n)}$ are obtained from the backbone network (convolutional neural network or Transformer) for the student and teacher networks at different levels of the intermediate feature layers, where n is the number of intermediate feature layers. Additionally, D_{GT} is the ground truth depth information, and $MF_{GT}^{x,y,z}$ is the motion field of the foreground target. When applying knowledge distillation technique to detection tasks, the issue of imbalanced information between foreground and background regions needs to be considered. Therefore, this method introduces a binary mask M based on the 2D bounding box $BBox_k$ to integrate the features of these two types of regions. The specific integration method is as follows:

$$M^k = \left\{ M_{i,j}^k \,|\, M_{i,j}^k = I\left[(i,j) \in BBox^k\right] \right\} \tag{1}$$

where i, j vary over the height and width of the feature map. If the feature map is foreground at i, j, then $M_{i,j}^k$ is 1, otherwise $M_{i,j}^k$ is 0.

Some works attempted to balance the use of depth-encoded positions with 3D embeddings and discretized grid coordinates. Thanks to knowledge distillation, our proposed method introduces semantic and motion features through a 3D positional encoding with depth and motion trends guided by the ground truth values of motion features. This method generates the 3D positional encoding (PE) under the guidance of D_{GT} and $MF_{GT}^{x,y,z}$ and incorporates PE into the feature self-attention computation of the student network. Firstly, a multi-layer perceptron (MLP) and bilinear interpolation are used on D_{GT} and $MF_{GT}^{x,y,z}$ to produce encoded depth features I_D and motion features I_{MF}, which have the same number of channels and match the size of the feature maps generated by the backbone network. Then, a cross-attention module is used to compute the cross-correlation map between the depth information and motion features, and they are fused into a unified depth-motion feature I_{DM}. Next, the argmax value of I_{DM} along the channel dimension is computed to obtain the mapping of depth position index, represented as P_{arg}. Then, the 3D embedding vector I_{emb} is generated using this index and the corresponding PE. Finally, I_{emb} is flattened after passing through the MLP to produce semantic query and key vectors with fused encoding of 3D position and motion features. The computation as mentioned above is as follows:

$$I_{emb} = P_{pe}\left(P_{arg}\left(MLP\left(I_{MF}\right)\right)\right) \tag{2}$$

To reduce computational burden, we use convolutional layers with kernel size 1 to build our MLP. Then, the query and key vectors interact through self-attention mechanism to generate the semantic feature I_{3d}^k incorporating depth and motion information. The distillation loss for the intermediate feature training process is as follows:

$$\begin{aligned} L_{inter} = &\sum_{k=1}^{n} \alpha_i \|M^k(I_{3d}^k - T^k)\|^2 \\ &+ \beta_i \|(1 - M^k)(I_{3d}^k - T^k)\|^2 \end{aligned} \tag{3}$$

where α_i and β_i are hyperparameters for foreground and background balance in the feature encoding part. The above is the part that introduces motion features in the teacher network training process.

After the motion feature encoding is completed, a decoder with knowledge distillation technique is needed to obtain the final network output in the subsequent process. For the 3D object detector based on encoder-decoder Transformer structure, the target feature vectors in the intermediate layers of the student and teacher network decoders are defined as $\{I_d^k\}_{k\in(1,\ldots,m)}$ and $\{T_d^k\}_{k\in(1,\ldots,m)}$, respectively, where m is the number of Transformer self-attention modules in the encoder. In order to balance foreground and background information, the Hungarian algorithm [25] is adopted to match the teacher network output with ground truth annotations, and then a foreground query mask M_f is generated for each target query vector of the feature layers. Similar to the above feature interaction method, this method utilizes cross-attention modules to guide the updated student network feature vectors I_{update}^k from the teacher network output. The overall interaction method is similar to the cross-attention module in the Transformer. The feature distillation part interacts teacher and student intermediate feature maps with a strategy like cross-attention mechanism, which is supervised by minimizing the Euclidean distance between teacher and student intermediate features. The loss function for this part is calculated as follows:

$$\begin{aligned} L_{distill} = &\sum_{k=1}^{m} \alpha_d \|M_f(I_t^k - T_d^k)\|^2 \\ &+ \beta_d \|(1 - M_f)(I_t^k - T_d^k)\|^2 \end{aligned} \tag{4}$$

where α_d and β_d are hyperparameters used for foreground and background balancing in the output decoding part. Feature maps with different resolution are fed into transformer blocks to calculate an attention map within the spatio-temporal domain. The well-trained transformer blocks have the ability to focus on the foreground, and thus the influence of the redundant background information is minimized.

The overall motion feature encoding and output decoding parts are trained end-to-end with four task-driven loss functions jointly supervised. These four loss functions are classification loss L_{cls}, 3D bounding box regression loss L_{reg},

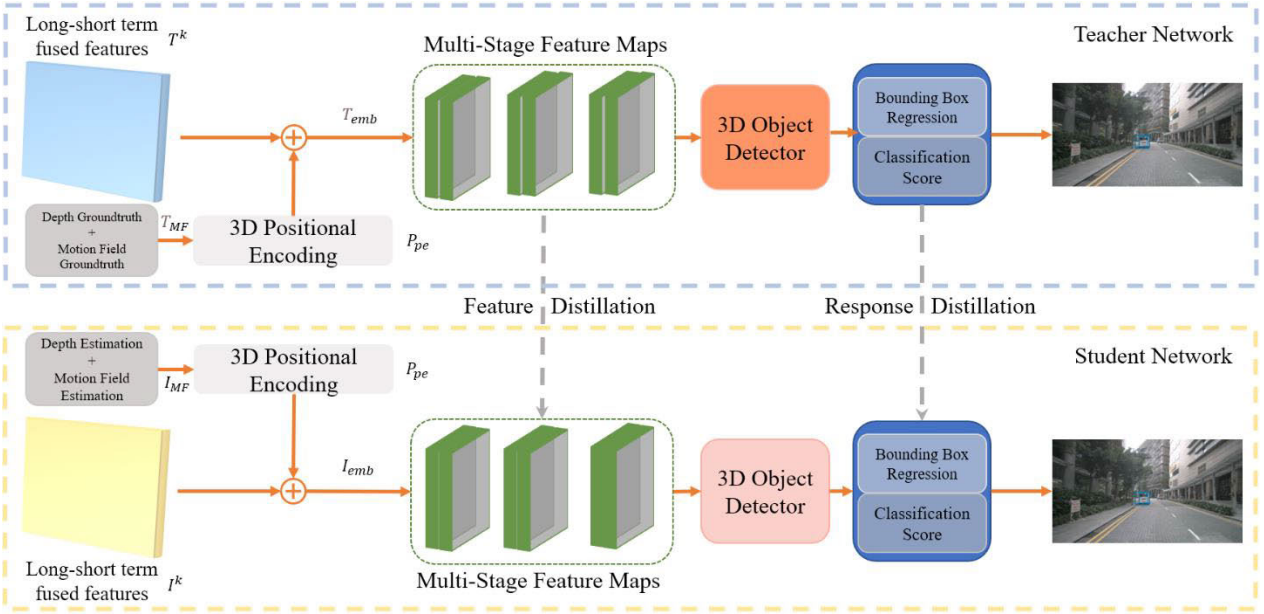

FIGURE 2. Overview of motion feature distillation.

depth loss L_d, and motion loss L_M. If the model adopts the results provided by the pre-trained depth estimation method, the depth loss will be set to zero during training. The overall loss function used for supervised training is as follows:

$$L_{total} = L_{cls} + L_{reg} + L_d + L_M + \alpha L_{inter} + \beta L_{distill} \quad (5)$$

where α and β are hyperparameters that control the balance of losses. Therefore, our proposed method is able to introduce fused depth and motion features in the feature encoding process, and utilize knowledge distillation techniques in the decoding process to give the decoder the ability to recognize and process the fused features. As a result, adaptability of our proposed model to the motion states of the observation points is improved. Accuracy and stability of 3D localization are enhanced.

IV. EXPERIMENTS AND DISCUSSIONS

A. IMPLEMENTATION DETAILS

Our proposed method was implemented using the programming language of Python. All the experiments were conducted on a PC with Nvidia A100 40G GPU and Intel I9 CPU, using PyTorch library version 1.8.1 with CUDA 11.3. The feature fusion part of our proposed method adopts BEVDepth as the baseline model and uses image features extracted from ResNet50 (pre-trained on the ImageNet dataset) as input. For the selection of the long-term and short-term spatio-temporal windows, the original image resolution is 1600×900. A time step of $T = 16$ was used for long-term fusion, while a step of $T = 2$ was used for short-term fusion. The previous BEV feature maps were saved during training and inference, and these historical feature maps were used in subsequent time steps to maintain high efficiency in the long-term fusion process. For the matching channels of short-term high-resolution image features, their feature dimensions were reduced to 64.

We evaluated our method on nuScenes dataset [26] which is a large-scale outdoor dataset focused on autonomous driving scenarios. It has diverse annotations to support all sorts of tasks. Each of the 40,157 annotated samples contains six monocular camera images with 360-degree field of view and a 32-beam LiDAR scan. Each camera has fixed setting of 12 Hz capture frequency and 2 Hz annotation rate.

The training method of multi-frame image motion feature distillation uses $\alpha_i = 1.0, \beta_i = 0.1$ for loss balancing at the feature encoding level and $\alpha_d = 1.0, \beta_d = 1.0$ as foreground and background balance hyperparameters in the decoder. AdamW optimizer was used for training with an initial learning rate of 0.0002 and a decay weight of 0.0001. The learning rate is decayed by 0.05 at each epoch, and the model is trained on the nuScenes dataset [26] with a batch size of 1 for 50 epochs.

The quantitative analysis was performed on the test set of the nuScenes dataset and compared with current state-of-the-art monocular 3D object detection methods in multiple metrics introduced in [26]. The ablation experiments were performed on the validation set of the nuScenes dataset to comprehensively analyze and validate the different components and parameters of the proposed method.

B. QUANTITATIVE RESULTS

Comparative analysis is carried out between the proposed monocular 3D object detection method and state-of-the-art methods on the nuScenes test set. Our algorithm is thoroughly trained on the complete data of the nuScenes training and validation sets. The complete results are shown in Table 1.

TABLE 1. Comparison of our proposed method with state-of-art methods on nuScenes test set.

nuScenes Test Set							
Method	$mAP \uparrow$	$NDS \uparrow$	$mATE \downarrow$	$mASE \downarrow$	$mAOE \downarrow$	$mAVE \downarrow$	$mAAE \downarrow$
FCOS3D [27]	0.358	0.428	0.690	0.249	0.452	1.434	0.124
DETR3D [28]	0.412	0.479	0.641	0.255	0.394	0.845	0.133
UVTR [29]	0.472	0.551	0.577	0.253	0.391	0.508	0.123
BEVFormer [14]	0.481	0.569	0.582	0.256	0.375	0.378	0.126
BEVDet4D [4]	0.451	0.569	0.511	0.241	0.386	0.301	0.121
PolarFormer [16]	0.493	0.572	0.556	0.256	0.364	0.439	0.127
PETRv2 [17]	0.512	0.592	0.547	0.242	0.360	0.367	0.126
BEVDepth [8]	0.520	0.609	0.445	0.243	0.352	0.347	0.127
BEVStereo [9]	0.525	0.610	0.431	0.246	0.358	0.357	0.138
Ours	0.560	0.652	0.395	0.225	0.342	0.225	0.122

The best results are highlighted in red, the second best results are highlighted in blue, and similarly for the following tables. The experimental results show that under relatively simple training settings and hardware requirements, our proposed method outperforms the current state-of-the-art BEV-based monocular 3D object detection algorithms in terms of 3D object detection performance indicators [26]. Compared with the current best-performing algorithm BEVStereo [9], our method can improve mAP by 6.7% and mean Average Translation Error (mATE) by 8.3%. This indicates that introducing more temporal information can effectively improve 3D localization accuracy. Low-resolution feature maps can provide stronger temporal information support while reducing computational complexity for temporal information further away in time. Similarly, after introducing the motion feature distillation method, the model's representation ability of temporal information becomes more intuitive, and the prediction of target motion in the scene becomes more accurate, as reflected in a significant improvement in mean Average Velocity Error (mAVE) by approximately 35.2%. With the introduction of long-term features from a larger temporal window, the motion distillation method can obtain richer motion information and model the movement of objects in 3D space more accurately. Therefore, the long short-term feature fusion and motion feature distillation proposed in this paper complement each other, and the combination of the two can effectively improve the performance of the monocular 3D object detection algorithm.

C. ABLATION STUDY

In order to analyze the impact of the long short-term feature fusion and motion feature distillation proposed in this paper on the overall performance of monocular 3D object detection, this section compares the proposed method with a baseline algorithm that does not use these two components. At the same time, detailed experimental verification and analysis of the parameter selection within each component were also conducted to demonstrate the effectiveness of each parameter.

1) EFFECTIVENESS OF LONG-SHORT SPATIO-TEMPORAL FUSION

The baseline model for the ablation experiments in this section is the BEVDepth model [8] without temporal information fusion. BEVDepth consists of the following parts:

i) depth estimation network: using encoded camera intrinsic and extrinsic parameters and a set of monocular images with non-overlapping viewpoints as input, outputting the depth value of each pixel in each image. The intrinsic and extrinsic parameters of the camera are provided by the dataset, including the camera's focal length, principal point, distortion coefficients, position, direction, and rotation matrix;

ii) depth correction network: using the output of the depth estimation network and the depth values obtained through annotated 3D object localization projections as input, outputting the corrected depth value. The purpose of the

TABLE 2. Effectiveness of different temporal windows on monocular 3D object detection.

nuScenes Validation Set				
Temporal window	$mAP \uparrow$	$NDS \uparrow$	$mATE \downarrow$	$mAVE \downarrow$
0	0.302	0.345	0.746	1.150
1	0.314	0.425	0.731	0.452
2	0.328	0.434	0.732	0.389
4	0.351	0.456	0.698	0.331
8	0.372	0.468	0.686	0.315
16	0.379	0.479	0.652	0.304
32	0.369	0.473	0.649	0.312

correction process is to reduce the errors introduced by the projection transformation and avoid additional burden on the subsequent 3D localization;

iii) rapid view transformation: using the corrected depth value and the camera intrinsic parameters as input, projecting the image feature from the camera view to BEV. To handle feature maps of different sizes and resolutions, this transformation aligns the features to the BEV space through rapid interpolation algorithms;

iv) multi-frame fusion module: using multi-frame historical BEV features as input, learning the correlation between different time frames through self-attention modules, and outputting the fused multi-frame enhanced BEV feature;

v) 3D object detection head: using the fused BEV feature as input, an additional regression branch is added to predict the height of the center point of the 3D bounding box in the BEV space, and finally outputting the 3D bounding box and confidence.

2) IMPACT OF TEMPORAL WINDOW SELECTION

Our proposed method replaces the fourth part of the above-mentioned baseline model with a long short-term spatio-temporal fusion strategy to improve the integration performance of temporal information. The baseline algorithm for experimental comparison is the BEVDepth model [8] without the fourth part. The ablation experiment compares the improvement of the overall model detection performance after adding low-resolution image features with longer temporal window, and the results are shown in Table 2. As can be seen from the table, although fusing a single time step can significantly reduce the speed prediction index mAVE by 60.7%, and increasing the NDS by 23.2%. The 3D spatial localization indicators mAP and mATE only have small improvements of 4.0% and 2.0%, respectively. With the selection of longer temporal windows in the experiment, more BEV features of multiple time steps are fused into the volume used for spatial localization BEV loss calculation, and the two metrics mAP and mATE representing spatial localization are improved by 20.7% and 10.8%, respectively between 1- and 16-time steps. This phenomenon indicates that compared with using a single time step to provide temporal information support, using BEV features of multiple time steps in a larger temporal window can significantly improve the potential for localization. However, the overall performance improvement reaches saturation when expanding the temporal window to 16-time steps, because the visible area of the scene beyond 16-time steps overlaps very little in the temporal domain, and the performance of object 3D localization cannot be further improved by computing disparities.

3) IMPACT OF DEPTH HYPOTHESIS SAMPLING METHODS

In Table 3, different depth hypothesis sampling methods were selected to compare their effects on optimizing short-term high-resolution temporal feature fusion, and to demonstrate the effectiveness of our proposed Top-k depth sampling method. The experiment starts with the baseline algorithm that uses a single frame, and without using depth hypothesis sampling method. The feature dimension in depth space is at least 112 for each corresponding pixel. Such computation will increase the running time by six times and significantly increase the usage of GPU memory. If 28 uniform sampling methods are used, the localization performance of the model is enhanced, but it will cause a 2.4 times speed reduction and an increase in GPU memory. When the depth sampling number is 7, the running speed of the algorithm can be improved, but the model's detection performance decreases. The Top-k sampling method selected in our proposed method outperforms the uniform sampling method with 28 sampling

TABLE 3. Effectiveness of different depth hypothesis sampling methods.

nuScenes Validation Set						
Methods	Sampling Points	FPS	Memory	$mAP \uparrow$	$NDS \uparrow$	$mATE \downarrow$
Baseline	-	17.2	3.3GB	0.302	0.345	0.746
Uniform	112	2.3	8.6GB	-	-	-
Uniform	28	6.8	4.1GB	0.352	0.379	0.690
Uniform	7	11.2	3.3GB	0.320	0.357	0.732
Top-k	7	11.8	3.3GB	0.341	0.385	0.678

TABLE 4. Effectiveness of short-term high-resolution and long-term low-resolution features.

nuScenes Validation Set					
Features Fusion	FPS	Memory	$mAP \uparrow$	$NDS \uparrow$	$mATE \downarrow$
Baseline	17.2	3.3GB	0.302	0.345	0.746
Short-term High-res	11.9	3.3GB	0.348	0.382	0.672
Long-term Low-res	15.6	3.7GB	0.385	0.474	0.654
Both	11.2	3.7GB	0.412	0.492	0.615

points in mATE, which indicates that selecting depth sampling points guided by monocular depth estimation prior can effectively improve the accuracy of 3D spatial localization.

4) IMPACT OF SHORT-TERM HIGH-RESOLUTION TEMPORAL FEATURE AND LONG-TERM LOW-RESOLUTION FEATURE

Table 4 shows a comparison of the 3D detection metrics and computational efficiency of the baseline algorithm with and without the addition of long-term low-resolution temporal features and short-term high-resolution temporal features. After adding short-term high-resolution temporal features, the computational speed decreased from 17.2 FPS to 11.9 FPS without any additional GPU memory consumption, while mATE is significantly improved by 9.9%. In comparison, the addition of long-term low-resolution temporal features significantly improves mAP by 27.5% with a small impact on computational speed (FPS is only decreased by 1.6). It is worth noting that the addition of only short-term high-resolution features decreases mATE by 9.9%, while the addition of long-term low-resolution features decreases it by 12.3%. This result indicates that both modules have a similar potential to enhance 3D localization capability. Finally, fusing both modules further improves the performance of all metrics, with an overall increase in NDS by 42.6%, demonstrating the complementary role of the two modules in 3D object detection.

5) EFFECTIVENESS OF MOTION FEATURE DISTILLATION

This section presents an ablation study on the effectiveness of the motion feature distillation method. Table 5 evaluates the effectiveness of the feature-level knowledge distillation in the motion feature distillation method, as well as the cross-attention distillation on the joint 3D PE and output decoder with motion features. Each component of the motion feature distillation method in Table 5 was tested on the nuScenes validation set.

Firstly, the teacher network trained with depth ground truth can accurately locate 3D objects, which is a prerequisite for the effectiveness of knowledge distillation. With only feature-level knowledge distillation (F), the teacher model's learned feature representation can be partially transferred to the student model, and adding the joint 3D PE with depth and motion features can further improve various metrics of 3D object detection. Among them, the mAVE is improved most significantly by 27.8%, while other metrics are improved by around 2-4%.

Secondly, after adding only the knowledge distillation module of the output decoder (O), the performance improvement effect is more significant than using only feature-level motion distillation. This may be because the main generation module for 3D object localization comes from the decoder, which can transfer the feature combination learned

TABLE 5. Effectiveness of motion feature distillation.

nuScenes Validation Set								
	F	PE	O	CA	$mAP \uparrow$	$NDS \uparrow$	$mATE \downarrow$	$mAVE \downarrow$
Teacher	Trained with Depth GT				0.554	0.598	0.454	0.154
Feature-level	√	-	-	-	0.412	0.502	0.628	0.327
	√	√	-	-	0.424	0.515	0.621	0.236
Output-level	-	-	√	-	0.408	0.491	0.632	0.308
	-	-	√	√	0.419	0.507	0.616	0.286
Both	√	-	√	-	0.432	0.525	0.593	0.258
	√	√	√	-	0.444	0.534	0.579	0.210
	√	-	√	√	0.453	0.541	0.573	0.239
	√	√	√	√	0.460	0.553	0.562	0.195

FIGURE 3. Visual results obtained by our proposed method on nuScenes validation set.

by the teacher model for localization to the student model after knowledge distillation. This is more effective for 3D localization with existing features. In addition, adding the cross-attention module (CA) can further improve overall performance. Finally, after adding all modules to the overall motion feature distillation method, the monocular 3D object detector can introduce motion features to enhance the accuracy of spatial depth judgment and better recombine fused features, thus obtaining superior 3D spatial localization capabilities.

TABLE 6. Comparison of efficiency and performance of our proposed method with state-of-art method.

nuScenes Validation Set					
Method	FPS	Memory	$mAP \uparrow$	$NDS \uparrow$	$mATE \downarrow$
BEVStereo [9]	1.8	4.8GB	0.372	0.500	0.598
Ours	11.2	3.7GB	0.425	0.529	0.573

6) EFFICIENCY IMPROVEMENT

Table 6 shows a comparison between our proposed method and another state-of-art monocular 3D object detection algorithm BEVStereo [9] in terms of both efficiency and performance. With the best setting, our proposed method can operate at the inference speed of 11.2 fps. The experimental results demonstrate that our method achieves better results on multiple metrics of 3D object detection while significantly improving the algorithm's efficiency. Compared to existing approaches that only fuse short-range high-resolution temporal features, our proposed model is simpler, more effective, and outperforms them in terms of performance.

D. QUALITATIVE RESULTS

Fig. 3 shows visual results obtained by our proposed method on the nuScenes validation set. The yellow 3D bounding boxes represent vehicles, while the blue ones represent pedestrians. The arrows at the center of each bounding box indicate the motion direction of the corresponding object. As can be seen from the figure, the proposed long short-term fusion method and motion feature distillation method achieve precise 3D localization of objects in the scene and accurate prediction of their 3D motion trends. Moreover, thanks to the introduction of temporal information and motion features, our proposed method is effective in detecting objects with certain occlusion. Therefore, introducing a larger temporal window is more effective for monocular 3D object localization. With the enhancement of more complex temporal information, the motion feature distillation method can capture more detailed temporal features of the object. In the case of occlusion, it can use motion features to infer the position of the object, further improving the accuracy of 3D object detection algorithm.

V. CONCLUSION AND FUTURE WORK

This paper proposes a monocular 3D object detection method based on long short-term feature fusion and motion feature distillation to better utilize the temporal disparity provided by the temporal information for more accurate spatial localization. The proposed method enhances the performance of monocular 3D object detection by first expanding the temporal window and using temporal features between multiple frames. Specifically, the proposed method improves the resolution of temporal neighboring feature maps to capture more detailed information, and reduces the resolution of feature maps for images far in time to reduce computational complexity and noise interference. A larger temporal window enables extraction of richer and more stable temporal features. To address the problem of temporal feature contamination caused by random motion of the observation point, we propose the motion feature distillation approach that uses motion features as a supervisory signal to assist in panoramic 3D depth estimation and improve the stability of 3D object localization. Experimental results show that the proposed method effectively uses temporal information to improve the accuracy and stability of monocular 3D object detection in panoramic environments. Our proposed method achieves a 6.7% improvement in mAP, 6.9% improvement in NDS, and 8.3% improvement in mATE on the nuScenes test set as compared to previous best art. Moreover, our method also achieves 35.2% improvement in predicting target velocities in the scene.

The domain transferability of 3D detection algorithms in autonomous driving environments is crucial for the safety of autonomous vehicles. Currently, vision-based 3D object detection methods show significant performance degradation on test data with large environmental changes. This is because vision-based 3D object detection relies heavily on the intra- and extra-parameters set by the image sensor dataset. The richness of the dataset can also affect the model's judgment of imaging conditions in different environments. Therefore, how to improve the model's generalization ability and reduce the deceptive influence of training data on the model's performance are the future research directions to enhance the applicability of vision-based 3D object detection algorithms.

REFERENCES

[1] Z. Xie, S. Zhou, M. Zheng, and F. Pei, "Research on self-supervised depth estimation algorithm of driving scene based on monocular vision," *Signal, Image Video Process.*, vol. 17, no. 4, pp. 991–999, Jun. 2023.
[2] H. Gao, D. Fang, J. Xiao, W. Hussain, and J. Y. Kim, "CAMRL: A joint method of channel attention and multidimensional regression loss for 3D object detection in automated vehicles," *IEEE Trans. Intell. Transp. Syst.*, early access, Nov. 10, 2022, doi: 10.1109/TITS.2022.3219474.
[3] D. Zhou, X. Song, J. Fang, Y. Dai, H. Li, and L. Zhang, "Context-aware 3D object detection from a single image in autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 18568–18580, Oct. 2022.
[4] J. Huang and G. Huang, "BEVDet4D: Exploit temporal cues in multi-camera 3D object detection," 2022, *arXiv:2203.17054*.
[5] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "MVSNet: Depth inference for unstructured multi-view stereo," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 767–783.
[6] G. Bae, I. Budvytis, and R. Cipolla, "Multi-view depth estimation by fusing single-view depth probability with multi-view geometry," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2842–2851.

[7] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D," in *Proc. ECCV*, in Lecture Notes in Computer Science, vol. 12359, 2020, pp. 194–210, doi: 10.1007/978-3-030-58568-6_12.

[8] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, and Z. Li, "BEVDepth: Acquisition of reliable depth for multi-view 3D object detection," 2022, *arXiv:2206.10092*.

[9] Y. Li, H. Bao, Z. Ge, J. Yang, J. Sun, and Z. Li, "BEVStereo: Enhancing depth estimation in multi-view 3D object detection with dynamic temporal stereo," 2022, *arXiv:2209.10248*.

[10] J. Park, C. Xu, S. Yang, K. Keutzer, K. Kitani, M. Tomizuka, and W. Zhan, "Time will tell: New outlooks and a baseline for temporal multi-view 3D object detection," 2022, *arXiv:2210.02443*.

[11] F. Jin, Y. Zhao, C. Wan, Y. Yuan, and S. Wang, "Unsupervised learning of depth from monocular videos using 3D–2D corresponding constraints," *Remote Sens.*, vol. 13, no. 9, p. 1764, May 2021.

[12] D. Zhao, C. Ji, and G. Liu, "Monocular 3D object detection based on pseudo multimodal information extraction and keypoint estimation," *Appl. Sci.*, vol. 13, no. 3, p. 1731, Jan. 2023.

[13] S. Qu, X. Yang, Y. Gao, and S. Liang, "MonoDCN: Monocular 3D object detection based on dynamic convolution," *PLoS ONE*, vol. 17, no. 10, Oct. 2022, Art. no. e0275438, doi: 10.1371/journal.pone.0275438.

[14] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *Proc. ECCV*, in Lecture Notes in Computer Science, vol. 13669, 2022, pp. 1–18, doi: 10.1007/978-3-031-20077-9_1.

[15] J. Huang, G. Huang, Z. Zhu, Y. Ye, and D. Du, "BEVDet: High-performance multi-camera 3D object detection in bird-eye-view," 2021, *arXiv:2112.11790*.

[16] Y. Jiang, L. Zhang, Z. Miao, X. Zhu, J. Gao, W. Hu, and Y.-G. Jiang, "PolarFormer: Multi-camera 3D object detection with polar transformer," 2022, *arXiv:2206.15398*.

[17] Y. Liu, J. Yan, F. Jia, S. Li, A. Gao, T. Wang, X. Zhang, and J. Sun, "PETRv2: A unified framework for 3D perception from multi-camera images," 2022, *arXiv:2206.01256*.

[18] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[19] K. Xu, L. Rui, Y. Li, and L. Gu, "Feature normalized knowledge distillation for image classification," in *Proc. ECCV*, in Lecture Notes in Computer Science, vol. 12370, 2020, pp. 664–680, doi: 10.1007/978-3-030-58595-2_40.

[20] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2599–2608.

[21] Y. Wang, X. Li, M. Shi, K. Xian, and Z. Cao, "Knowledge distillation for fast and accurate monocular depth estimation on mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 2457–2465.

[22] X. Guo, S. Shi, X. Wang, and H. Li, "LIGA-stereo: Learning LiDAR geometry aware representations for stereo-based 3D detector," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3133–3143.

[23] Z. Chong, X. Ma, H. Zhang, Y. Yue, H. Li, Z. Wang, and W. Ouyang, "MonoDistill: Learning spatial features for monocular 3D object detection," 2022, *arXiv:2201.10830*.

[24] J. Yang, S. Shi, R. Ding, Z. Wang, and X. Qi, "Towards efficient 3D object detection with knowledge distillation," 2022, *arXiv:2205.15156*.

[25] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Res. Logistics Quart.*, vol. 2, nos. 1–2, pp. 83–97, Mar. 1955.

[26] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11621–11631.

[27] T. Wang, X. Zhu, J. Pang, and D. Lin, "FCOS3D: Fully convolutional one-stage monocular 3D object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 913–922.

[28] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "DETR3D: 3D object detection from multi-view images via 3D-to-2D queries," in *Proc. Conf. Robot Learn.*, vol. 164, 2022, pp. 180–191.

[29] Y. Li, Y. Chen, X. Qi, Z. Li, J. Sun, and J. Jia, "Unifying voxel-based representation with transformer for 3D object detection," 2022, *arXiv:2206.00630*.

HENAN HU received the B.Eng. degree in mechanical engineering from the Changchun University of Science and Technology, China, in 2018. She is currently pursuing the Ph.D. degree with the University of Chinese Academy of Sciences, China. Her research interests include 3D object detection and deep learning.

MUYU LI received the B.Sc. degree in electrical science and technology from the University of Science and Technology of China, in 2014, and the Ph.D. degree from the University of Chinese Academy of Sciences, China. Currently, he is with the Center for Intelligent Multidimensional Data Analysis Ltd., Hong Kong. His research interests include visual tracking and deep learning.

MING ZHU is currently a Research Fellow and a Supervisor of Ph.D. candidates with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences. His research interests include digital image processing, television tracking, and automatic target recognition technology.

WEN GAO is currently an Associate Research Fellow with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences. Her research interests include digital image processing, television tracking, and automatic target recognition technology.

PEIYU LIU received the master's degree from Jilin University. She is currently an Engineer with the Shenyang Aircraft Design & Research Institute. Her research interest includes avionics system design.

KWOK-LEUNG CHAN received the M.Sc. degree from the Institute of Science and Technology, University of Wales, U.K., and the Ph.D. degree from the College of Medicine, University of Wales. He is currently an Assistant Professor with the Department of Electrical Engineering, City University of Hong Kong. His research interests include image processing and deep learning.