Focus on your Geometry: Exploiting the Potential of Multi-Frame Stereo Depth Estimation Pre-training for 3D Object Detection

Zichen Wang, Zhuokun Yao, Jianwei Zhang, Ye Zheng, Zhengyuan Zhang,
Shuang Deng, Yajing Liu and Hao Liu
Autonomous Driving Department of X Division, JD Logistics, Beijing, China
{wangzichen, yaozhuokun1, zhangjianwei64, zhengye12, zhangzhengyuan5, dengshuang10, liuyajing25, liuhao163}@jd.com

2024 International Joint Conference on Neural Networks (IJCNN) | 979-8-3503-5931-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/IJCNN60899.2024.10650924

***Abstract*—Existing camera-based 3D object detection methods yield inaccurate position, scale, and orientation results due to the inherent challenge of ill-posed depth estimation from 2D images. Recent research has demonstrated that pre-training depth estimation from a single frame substantially enhances the quality of camera-based 3D object detection. We hypothesize that integrating multi-view stereo matching technology into the pre-training process can equip the backbone model with superior geometric feature extraction capabilities, thereby further improving 3D object detection performance. Building upon this premise, we propose MVS3D, a novel depth estimation pre-training method for camera-based 3D object detection. MVS3D incorporates a VMS (Video-stream-based Multi-view Stereo) module and a PME (Pose and Motion Estimation) module, which collectively encourage the backbone to explicitly learning 3D geometric information from image streams through stereo matching. Our method enables existing camera-based 3D object detection frameworks to seamlessly integrate our pre-trained backbone weight, thereby enhancing detection performance without necessitating extensive modifications. Extensive experimental results on nuScenes dataset show that loading the pre-trained weight from MVS3D can significantly improve the mean average precision (mAP) and nuScenes detection score (NDS) of both existing single-frame and multi-frame camera-based methods.**

I. Introduction

LIDAR and cameras are the two principal sensors utilized in the realm of autonomous driving for the perception of 3D objects. Although LIDAR-based methodologies [1]–[5] are typically associated with higher precision in determining 3D positions and orientations, camera-based techniques [6]–[21] have attracted significant interest due to their cost-effectiveness, rich semantic information, and superior capability in long-range object detection. For a significant duration, a notable discrepancy in recognition accuracy has been observed between camera-based 3D object detectors and point cloud-based methods, a divide stemming from the intricate challenges of perspective transformation and the inherently difficult task of depth estimation. This divide is intrinsically connected with geometric information such as position, scale, and orientation.

Historically, several approaches [9] have been undertaken to secure a pre-trained model by directly fitting the depth of a single 2D image from extensive datasets, with the goal of augmenting the geometric representational capacity of the backbone network tasked with fundamental feature extraction. This method of pre-training has yielded promising results. However, directly regressing depth from a single 2D image is tantamount to an overfitting exercise. Such a method not only lacks geometric validity, but also demonstrates fragility in response to variations in camera intrinsics and extrinsics, thereby constraining the potential of camera-based 3D object detection techniques. So the question arises naturally: can we exploit the capabilities of depth estimation pre-training to further improve the performance of camera-based 3D object detection? The answer is affirmative.

Intuitively, the conundrum of ill-posed depth estimation can be considerably alleviated by employing multi-view stereo (MVS) techniques [22]–[24]. MVS leverages multiple images from various viewpoints to facilitate robust matching and deduce depth. Autonomous vehicles, outfitted with cameras, inherently produce video streams in sequence with consistent triggering and exposure intervals, rendering them perfectly suited for MVS-based depth estimation. By crafting a multi-view setup from these sequential frames, we can more effectively deploy MVS methodologies. Based on that, by incorporating pose transformation constraints across successive frames, we can refine the feature extraction network's proficiency in capturing geometrically informative features. Such enriched features can markedly elevate the efficacy of camera-based 3D object detection methods by supplying more robust geometric priors and enhancing their detection capabilities.

Base on above discussions, we propose a novel multi-frame stereo-matching-based 3D object detection pre-training method, named MVS3D, focusing on improving the geometric representation ability of the pre-trained model to improve 3D object detection performance. Within MVS3D, the introduced VMS module conducts temporal stereo matching during training, significantly enhancing the accuracy of depth predictions and compelling the backbone to explicitly assimilate stereoscopic information, thereby focusing on the intrinsic geometric relationships within the scene rather than merely overfitting to depth perception. We then present the PME module, which predicts the transformation matrices between the key frame and each historical frame, further endowing the backbone with a more potent stereo representation by explicitly mining the 3D

geometric features. Once the pre-training phase is complete, the backbone weights, now imbued with an enhanced capacity for geometric representation, can be flawlessly integrated as a pre-training step in most extant camera-based 3D object detection frameworks to further amplify their performance.

Extensive experiments conducted on the nuScenes benchmark [25] demonstrate that integrating our pre-trained model into existing CNN-based [7], [8], attention-based [14], [19], and BEV-based [10], [11], [16] methodologies yields consistent enhancements over models pre-trained on ImageNet [26] or via traditional single-image depth pre-training [9]. In summary, the main contributions of this work are threefold:

- We propose a novel multi-frame stereo-matching-based 3D object detection pre-training method, which incorporates the VMS module to enable the backbone network to explicitly extract 3D geometric information during training.
- We introduce the PME module, which predicts the inter-frame pose transformations, thereby facilitating the feature extraction network to acquire features with enhanced geometric representational capabilities.
- Our experimental findings demonstrate that by utilizing our pre-trained backbone, existing camera-based 3D object detection frameworks can achieve substantial improvements in detection performance.

II. RELATED WORK

In this section, we first briefly divide the existing 3D object detection methods into two categories, *i.e.*, single-frame approaches [6]–[15] and video-based multi-frame approaches [16]–[21], and introduce their core ideas, flaws and admirable points respectively. Then we introduce representative works that utilize stereo matching technology to enhance depth estimation capabilities [22], [23], [27]–[31].

A. Single-frame 3D Object Detection Methods

An early method for monocular 3D object detection from single-view images is Mono3D [32], which utilizes geometric priors to achieve the 2D-3D projection. Then many works have made their effort on extending 2D detection approaches to 3D task [6]–[8]. DD3D [9] further explores the potential of single-image depth estimation pre-training on the large-scale datasets, and yields more accurate positions of 3D objects. The above methods usually adopt CNN-based architectures to process single-view images, failed to improve detection performance through utilizing multi-view overlaps. To better process the surround-view detection task, mainstream methods are divided into two categories, *i.e.*, sparse attention-based detectors [13]–[15] and dense BEV-based [10]–[12] detectors. Sparse methods are represented by PETR [14], which designs the inspiring 3D position embedding strategy and employs sparse queries to perform attention-based perception. Focal-PETR [15] further limits the scope of global attention to the foreground area, achieving dual improvements in accuracy and computing efficiency. BEV-based methods transform image features into bird-eye-view space through *Lift, Splat, Shoot (LSS)* [33] or attention layers. BEVDet [11] and BEVDepth [10] predict depth maps for multi-view images, and build dense BEV feature maps by voxel-pooling [34], which is a high-performance implementation of LSS. BEVFormer [12] constructs grid-shaped queries in BEV space, and aggregates image features based on the attention mechanism [35].

B. Multi-frame 3D Object Detection Methods

Unlike single-frame methods, multi-frame approaches [16]–[21] extract features from the image stream and fuse them to perceive the key-frame target objects. The fused temporal features contain rich motion information, which can greatly improve the location and velocity estimation performance, leading to stable and accurate 3D object detection. StreamPETR [21] and Sparse4Dv2 [20] are the representative works extending single-frame sparse attention-based detectors to multi-frame video-based detectors, which design ingenious motion fusion strategies to propagate the queries of historical frames forward. To effectively explore the temporal information and improve perception performance, multi-frame BEV-based detectors [16]–[18] fuse the 4D BEV features through explicit warping alignment. BEVDet4D [16] lifts the scalable BEVDet paradigm from the spatial-only 3D space to the spatial-temporal 4D space. VideoBEV [18] proposes a RNN-style temporal fusion approach for camera-based BEV 3D detection. SOLOFusion [17] balances low-resolution, long-term and high-resolution, short-term temporal fusion to maximize camera-only depth estimation potential with high efficiency. Although the above-mentioned multi-frame methods significantly improve the perception accuracy and stability by exploring inter-frame motion information, they still suffer from the ill-posed single-image depth estimation without video-base stereo matching.

C. Stereo-matching-based Depth Estimation

As mentioned before, despite considerable progress, the bottleneck of the existing 3D object detection methods is the inability to solve the depth estimation problem well. Predicting depth from a single image is obviously ill-posed, therefore, many multi-view depth estimation approaches [22], [23], [27]–[31] are proposed. Based on multi-view stereo matching, many algorithms [22] construct cost volume to predict high-quality depth maps. MVSNet [27] is the first learning-based model to apply stereo matching to depth estimation, and many works derived from it [28]–[31] have been proposed to optimize its computational efficiency and performance.

In this paper, we incorporate stereo matching depth estimation into 3D object detection task, and surprisingly discover that the backbone obtains the ability to explicitly dig geometric features in this way, leading to significant improvements in the performance of both single-frame and multi-frame 3D object detection methods when loading our pre-trained backbone.

III. METHODOLOGY

In this section, we delineate the architecture of MVS3D, a pre-trained model architected to augment 3D object detection

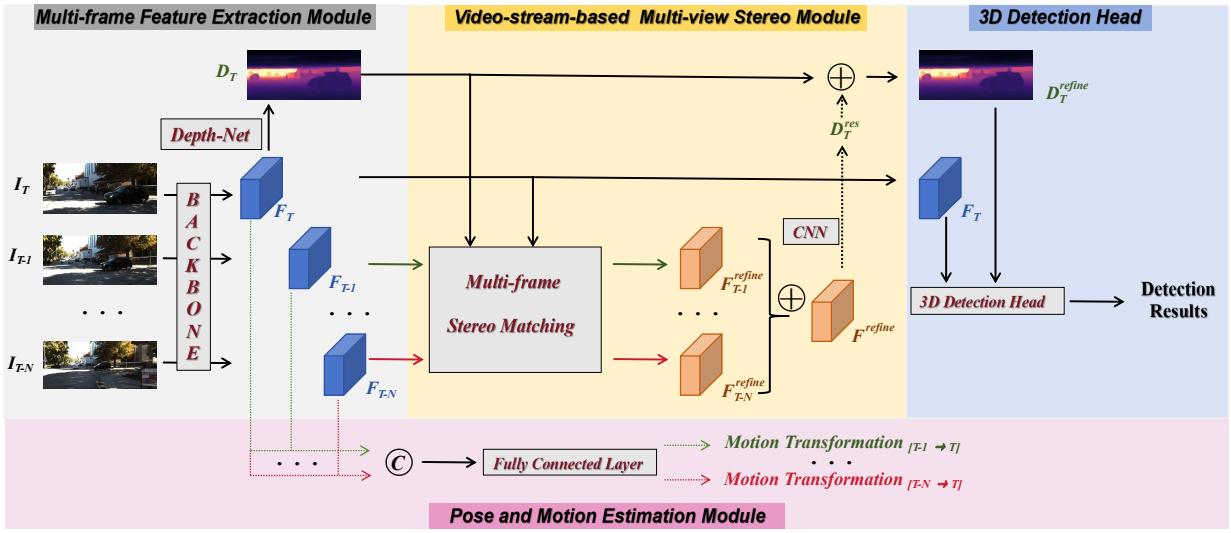

Fig. 1. Illustration of MVS3D. The *Multi-frame Feature Extraction Module* extracts features of all frames through a shared-weight backbone, and estimates the rough depth map of key frame. The VMS module performs stereo matching from key frame to each historical frame, and refines the predicted depth map. The PME module predicts the inter-frame transformation matrices.

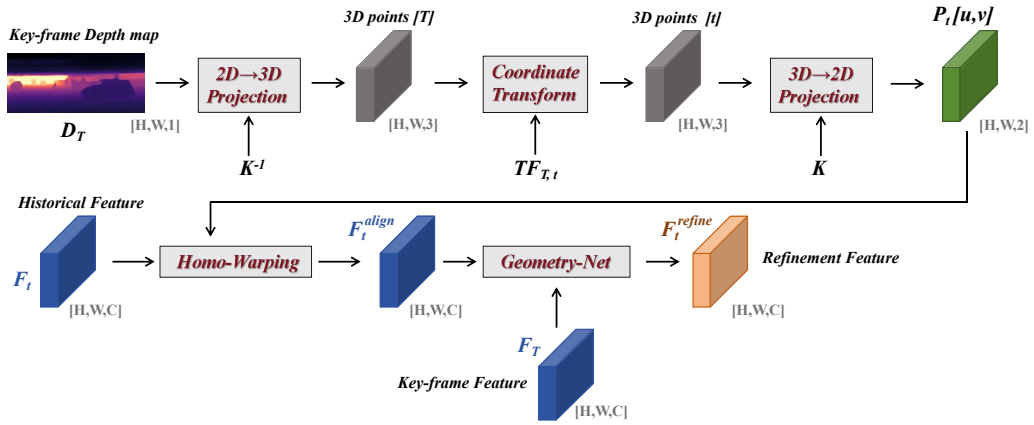

Fig. 2. Detailed illustration of *Multi-frame Stereo Matching* Module. The image coordinate mapping ($P_t[u,v]$) from key frame to each historical frame is calculated by Eq (4), which is used to warp the historical feature. The aligned historical feature will be fed into the Geometry-Net together with the key-frame feature to explore the rich geometric information which contains valuable clues for depth estimation refinement.

capabilities for both single-frame and multi-frame camera-based systems. MVS3D harnesses multi-view stereo matching technology to extract superior geometric features. Figure 1 illustrates the four main modules of MVS3D: the multi-frame feature extraction module, the video-stream-based multi-view stereo (VMS) module, a conventional 3D object detection head, and an additional pose and motion estimation (PME) module.

A. Multi-frame Feature Extraction Module

For training, MVS3D ingests a series of images $\{I_t \in \mathbb{R}^{H\times W\times 3} \mid T-N \le t \le T\}$, where t indexes the temporal sequence, with I_T denoting the current key frame and $\{I_t \mid T-N \le t \le T-1\}$ representing the preceding frames. Each frame is accompanied by a transformation matrix $TF_t \in \mathbb{R}^{4\times 4}$, which relates the ego-vehicle's coordinate system to a global frame of reference. The transformation matrix between frames is computed as:

$$TF_{t1,t2} = TF_{t2}^{-1} \times TF_{t1}. \tag{1}$$

A shared-weight neural backbone network is employed to extract features across all frames, as per the following formulation:

$$\{F_t \in \mathbb{R}^{H_d\times W_d\times C} \mid T-N \le t \le T\} = \{Backbone(I_t)\} \tag{2}$$

where H_d and W_d denote the height and width of the down-sampled feature maps, respectively. Subsequently, the feature map of the current frame F_T is fed into the Depth-Net which consists of several cascading convolution layers to predict a dense depth map of the scene:

$$D_T \in \mathbb{R}^{H_d\times W_d} = DepthNet(F_T). \tag{3}$$

As previously discussed in Section I, traditional depth estimation from 2D images is prone to overfitting and generalization issues. To mitigate this, we propose the VMS module, which explicitly exploits the 3D geometric structure of the scene.

B. Video-stream-based Multi-view Stereo Module

The VMS module refines the depth map by applying multi-frame features $\{F_t | T - N \leq t \leq T\}$ and the preliminary key-frame depth estimation D_T through stereo matching. As shown in Fig. 2, the process begins by computing the image coordinate mapping $P_{T_n}[u, v] \in \mathbb{R}^{H_d \times W_d \times 2}$ from the key frame ($t = T$) to each historical frame ($t = T_n$) using the equation:

$$P'_{T_n}[uz, vz, z] = K \times TF_{T,T_n} \times K^{-1} \times (D_T \cdot P'_T[u, v, 1]),$$
$$P_{T_n}[u, v] = P'_{T_n}[..., 0:2]/P'_{T_n}[..., 2:3], \tag{4}$$

where K is the intrinsic camera matrix, and u,v represent the pixel coordinates. The historical feature maps $\{F_t | T - N \leq t \leq T - 1\}$ are then warped to align with the key-frame coordinates. The aligned features F_t^{align} are combined with the key feature F_T and processed by Geometry-Net, which refines the depth map by leveraging the geometric correspondence between them. Geometry-Net is a stack of convolutional layers.

Contrary to traditional stereo matching that depends on similarity metrics for depth map validation [36], our Geometry-Net autonomously learns to extract geometric features, enabling the backbone to explicitly infer 3D geometric information that is critical for accurate depth estimation. The cumulative refinement features produce a depth residual map, which is added to the initial depth map, resulting in a refined depth estimation. This refined depth map D_T^{refine}, together with the key-frame feature F_T, is input into the 3D detection head.

The introduction of the stereo matching mechanism allows the MVS3D backbone to more effectively extract 3D geometric information compared to conventional methods that predict depth from single images. Existing methods can readily integrate MVS3D's pre-trained backbone, benefiting from its enhanced geometric feature extraction capabilities, which leads to more precise depth estimation and 3D object detection.

C. Pose and Motion Estimation Module

To further emphasize the extraction of 3D geometric features, we incorporate a PME module that predicts the transformation matrices between the key frame and each historical frame. This module encourages the backbone to concentrate on stereo information, thereby improving its representational strength and generalization capacity.

D. Loss Function

Training of MVS3D is supervised by three signals: the ground truth for target 3D objects $\{obj_i^{GT} | 1 \leq i \leq M\}$, the ground truth for the key-frame depth map D_T^{GT}, and the ground truth for the motion transformation matrices $\{TF_{t,T}^{GT} | T - N \leq t \leq T - 1\}$, with M indicating the total number of target objects. The composite loss function is expressed as:

$$L = L_{3D} + 0.5 \times L_{depth} + 0.5 \times L_{TF},$$
$$L_{depth} = L1(D_T^{GT}, D_T) + L1(D_T^{GT}, D_T^{refine}),$$
$$L_{TF} = \frac{1}{N} \sum L1(TF_{t,T}^{GT}, TF'_{t,T}), \tag{5}$$

where L_{3D} is consistent with traditional 3D object detection approaches [8], [11], [14], incorporating a focal loss [37] for classification and an L1 loss for bounding box regression. The ground truth transformation matrix $TF_{t,T}^{GT}$ is derived using Equation (1), and $TF'_{t,T}$ is the output of the PME module.

IV. EXPERIMENT

In this section, we first explain the dataset and evaluation metrics adopted in our experiments. Then, we provide a detailed introduction to the implementation settings. Finally, to further demonstrate the universality of our method, we conduct multiple sets of comparative experiments on various mainstream methods.

A. Dataset and Evaluation Metrics

We evaluate all of our experiments on nuScenes [25], a large-scale autonomous driving benchmark with 700, 150 and 150 scenarios for training, validation and testing, respectively, which are captured from multiple onboard sensors including six cameras, one LIDAR, and five radar. Each sequence is roughly 20s long, with a sampling rate of 20 frames/second. For camera-based 3D detection task, we report the standard NDS (nuScenes Detection Score) as the key indicator, which is a consolidated scalar metric capturing all aspects of the detection task including mAP(mean Average Precision) and five TP (true positive) metrics, *i.e.*, mATE (mean Average Translation Error), mASE (mean Average Scale Error), mAOE (mean Average Orientation Error), mAVE (mean Average Velocity Error) and mAAE (mean Average Attribute Error), via

$$NDS = \frac{1}{10}[5mAP + \sum_{mTP}(1 - min(1, mTP))]. \tag{6}$$

B. Implementation Details

Our MVS3D is implemented in PyTorch. We train our model using a batch size of 32 for 24 epochs, which takes about 50 hours using two NVIDIA A100 GPUs. Input images are randomly cropped and scaled to 512×1408. Adam optimizer [38] is used to optimize our network with default parameters ($\beta_1 = 0.9$ and $\beta_2 = 0.999$). We use the representative ResNet-50 [39] as the image backbone to train our MVS3D and conduct experiments. For the depth supervision signal used in the training process, we project the point cloud data in nuScenes from LIDAR coordinate to image plane utilizing camera intrinsics and extrinsics, without introducing external depth estimation dataset.

TABLE I
QUANTITATIVE RESULTS OF STATE-OF-THE-ART 3D OBJECT DETECTORS ON THE NUSCENES VALIDATION DATASET. WE HIGHLIGHT THE BEST RESULTS IN **BOLD**.

Architecture	Method	Pre-training	Resolution	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓	NDS↑
CNN-based	PGD [7]	R50-ImageNet	900 × 1600	0.2848	0.8024	0.2718	0.5326	**1.2236**	**0.1765**	0.3640
		R50-MVS3D	900 × 1600	**0.3334**	**0.7533**	**0.2650**	**0.5254**	1.3116	0.1887	**0.3935**
	FCOS3D [8]	R50-ImageNet	512 × 1408	0.2701	0.8272	**0.2649**	**0.5091**	1.1562	0.1680	0.3581
		R50-MVS3D	512 × 1408	**0.3170**	**0.7974**	0.2662	0.5099	**1.1430**	**0.1557**	**0.3856**
Attention-based	PETR [14]	R50-ImageNet	512 × 1408	0.3174	0.8397	**0.2796**	0.6158	0.9543	0.2326	0.3665
		R50-MVS3D	512 × 1408	**0.3683**	**0.7934**	0.2881	**0.6120**	**0.8812**	**0.2305**	**0.4036**
	PETRv2 [19]	R50-ImageNet	256 × 704	0.3490	0.7000	**0.2750**	0.5800	0.4370	**0.1870**	0.4560
		R50-MVS3D	256 × 704	**0.3682**	**0.6808**	0.2813	**0.5777**	**0.4301**	0.1931	**0.4678**
BEV-based	BEVDet [11]	R50-ImageNet	256 × 704	0.2828	0.7734	0.2884	0.6976	0.8637	0.2908	0.3500
		R50-MVS3D	256 × 704	**0.3096**	**0.7564**	**0.2818**	**0.6448**	**0.8087**	**0.2756**	**0.3781**
	BEVDepth [10]	R50-ImageNet	256 × 704	0.3304	0.7021	0.2795	**0.5346**	0.5530	0.2274	0.4355
		R50-MVS3D	256 × 704	**0.3559**	**0.6717**	**0.2731**	0.5442	**0.4818**	**0.2212**	**0.4588**
	BEVDet4D [16]	R50-ImageNet	256 × 704	0.3139	0.6908	**0.2818**	0.5492	0.3809	**0.1963**	0.4470
		R50-MVS3D	256 × 704	**0.3381**	**0.6898**	0.2856	**0.5255**	**0.3311**	0.2067	**0.4652**

TABLE II
ABLATION EXPERIMENTAL RESULTS ON THE NUSCENES VALIDATION DATASET. WE HIGHLIGHT THE BEST RESULTS IN **BOLD**. SDE DENOTES THE CONVENTIONAL SINGLE-FRAME DEPTH ESTIMATION PRE-TRAINING.

Method	ImageNet	SDE	VMS	PME	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓	NDS↑
FCOS3D [8]	✓				0.2701	0.8272	**0.2649**	**0.5091**	1.1562	0.1680	0.3581
	✓	✓			0.2913	0.8192	0.2711	0.5127	1.1533	0.1669	0.3687
	✓		✓		0.3101	0.8025	0.2703	0.5108	1.1536	**0.1502**	0.3817
	✓		✓	✓	**0.3170**	**0.7974**	0.2662	0.5099	**1.1430**	0.1557	**0.3856**

TABLE III
THE EFFECTIVENESS OF STEREO MATCHING IN MULTI-FRAME PRE-TRAINING. WE CONDUCT EXPRIMENTS ON THE NUSCENES VALIDATION DATASET. WE HIGHLIGHT THE BEST RESULTS IN **BOLD**.

Method	Pre-trained Backbone	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓	NDS↑
PETR [14]	R50-Stream-PETR	0.3251	0.8105	**0.2773**	0.6201	0.9259	0.2318	0.3760
	R50-MVS3D	**0.3683**	**0.7934**	0.2881	**0.6120**	**0.8812**	**0.2305**	**0.4036**
PETRv2 [19]	R50-Stream-PETR	0.3417	0.6984	**0.2779**	0.5819	0.4413	**0.1899**	0.4519
	R50-MVS3D	**0.3682**	**0.6808**	0.2813	**0.5777**	**0.4301**	0.1931	**0.4678**

C. Experimental Results Analysis

To prove the effectiveness of MVS3D, we conduct extensive experiments on extant state-of-the-art methods on the nuScenes val set [25]. The comparative trials covers various mainstream network architectures including CNN-based methods (PGD [7] and FCOS3D [8]), sparse attention-based approaches (PETR [14] and PETRv2 [19]) and dense BEV-based algorithms (BEVDet [11], BEVDepth [10] and BEVDet4D [16]), in which PETRv2 and BEVDet4D are video-based multi-frame detectors and the others are single-frame frameworks. For fairness, we only load the backbone weights of MVS3D during model initialization, leaving other training parameters such as image resolution and training epochs unchanged.

1) Quantitative Analysis: All the above existing single-frame and multi-frame detectors directly regress depth from a single 2D image, which is an ill-posed problem severely hindering their high-quality camera-based perception. In Table I, we carry out comparative experiments on existing approaches to verify the effectiveness of our pre-trained backbone. First, among typical monocular CNN-based detectors, PGD [7] and FCOS3D [8] achieve the NDS improvements of 2.95% and 2.75% respectively, with general promotions in mAP (4.86% and 4.69% respectively) and five TP metrics. It is worth noting that MVS3D-pre-trained FCOS3D [8] outperforms the native-pre-trained PETR [14], which is generally regarded as a powerful attention-based multi-view detector than CNN-based monocular detectors, comprehensively showing that

Fig. 3. Visualization of the detection results of MVS3D-pre-trained FCOS3D on the nuScenes test dataset.

MVS3D endows the backbone with a more potent geometric representation. Then we perform experimental analysis on the single-frame attention-based vision detector PETR [14]. Since the original PETR [14] directly perceives the target 3D objects without explicit supervision of depth ground truth during training, loading our MVS3D-pre-trained backbone brings it the most significant performance improvement in all trials, with the 5.09% mAP and 3.71% NDS increase. This considerable enhancement of perception capacity is attributed to our geometry-aware stereo depth estimation pre-training. Moreover, it is gratifying that even for BEV-based detectors that have explicitly deduced and supervised the dense depth distribution during training, MVS3D still supplies an further enhanced geometric representational capacity for the backbone, resulting in the 2.81% and 2.33% promotion in NDS for BEVDet [11] and BEVDepth [10] respectively. This improvement in depth-supervised BEV-based methods confirms the fact that securing a pre-trained model that directly fits the depth without stereo matching constrains the potential of digging stereoscopic information, and can be further improved by compelling the backbone to explicitly explore geometric features with rich stereo clues. Finally, we also perform experiments on two representative multi-frame frameworks, *i.e.*, attention-based PETRv2 [19] and BEV-based BEVDet4D [16], surprisingly find that MVS3D pre-training still works for them. The mAP and NDS of PETRv2 increase by 1.92% and 1.18% respectively, and MVS3D-pre-trained BEVDet4D surpasses the native-pre-trained model on mAP and NDS by 2.42% and 1.82% respectively. Although these video-based detectors perform temporal feature fusion to explore the inter-frame motion transformation, they still lack the explicit geometry-aware feature extraction, so that MVS3D can still markedly elevate their vision perception. Overall, integrating our MVS3D-pre-trained backbone yields consistent enhancements for existing state-of-the-art 3D object detectors under various architectures, whether single-frame or video-based approaches, with or without depth supervision during training.

2) Qualitative Analysis: To prove the effectiveness of our proposed MVS3D, we visualize the detection results of MVS3D-pre-trained FCOS3D [8] in Fig. 3. It shows that the single-frame CNN-based monocular detector can also deliver trustworthy detection results through loading our geometry-aware backbone.

D. Ablation Studies

We verify the effectiveness of our VMS module and PME module in Table II. The baseline method is FCOS3D [8] pre-trained on ImageNet [26], which achieves 27.01% mAP and 35.81% NDS. Since VMS module uses additional point cloud data to supervise the depth prediction compared with the native ImageNet pre-training, we set up a comparative experiment of conventional SDE (single-image depth estimation) pre-training, as shown in the second row of Table II, which performs single-image depth prediction without stereo matching. Consistent with the conclusion of previous works [9], SDE makes the backbone more proficient in depth perception, leading to a 1.06% improvement in NDS. Despite progress, directly deducing depth from a single 2D image lacks geometric validity, and can be further enhanced through multi-frame stereo matching. As shown in the second row of Table II, our VMS pre-training significantly outperforms SDE under the same training data and supervision, fully demonstrating that the VMS module imbues backbone with an improved capability for geometric representation. In the last row of the table, we further verify the effectiveness of our PME module, which enables the backbone to further assimilate stereoscopic information and achieves the best increase in mAP (4.69% $\uparrow$) and NDS (2.75% $\uparrow$).

E. Effectiveness of Stereo Matching

In addition to introducing stereo matching technology, MVS3D also performs multi-frame temporal fusion during training, both of which can improve the representation and generalization of backbone to some extent. We point out that

the powerful performance of MVS3D-pre-trained backbone mainly comes from stereo matching technology rather than temporal feature fusion. To verify our point, we compare MVS3D-pre-trained PETR [14] and PETRv2 [19] with loading weights from StreamPETR [21], which is a video-based detector fusing temporal features. As shown in Table III, the proposed MVS3D achieves higher mAP (4.32% ↑ and 2.65% ↑ respectively) and NDS (2.76% ↑ and 1.59% ↑ respectively) on both downstream 3D detection models, further verifying the effectiveness of the VMS module and PME module. It is worth noting that we only perform stereo matching during training and directly load the pre-trained model for the downstream 3D detection task, which means that the model can guarantee very fast inference speed, just like the single-frame algorithms.

V. Conclusion

In this paper, we exploit the potential of multi-frame stereo depth estimation pre-training for existing camera-based 3D object detectors. The introduced VMS and PME module enhance the representation and generalization of backbone by compelling the model to explicitly assimilate geometric information, significantly alleviating the conundrum of ill-posed depth estimation. Extensive experiments shows that loading our MVS3D-pre-trained backbone considerably promotes the performance of extant single-frame and multi-frame 3D object detection frameworks, providing an inspiring idea to achieve higher-quality 3D detection pre-training.

References

[1] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 697–12 705.

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

[3] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2017, pp. 1907–1915.

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

[5] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2018, pp. 4490–4499.

[6] Z. Liu, Z. Wu, and R. Tóth, "Smoke: Single-stage monocular 3d object detection via keypoint estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 996–997.

[7] T. Wang, X. Zhu, J. Pang, and D. Lin, "Probabilistic and Geometric Depth: Detecting objects in perspective," in *Conference on Robot Learning (CoRL) 2021*, 2021.

[8] ——, "Fcos3d: Fully convolutional one-stage monocular 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 913–922.

[9] D. Park, R. Ambrus, V. Guizilini, J. Li, and A. Gaidon, "Is pseudo-lidar needed for monocular 3d object detection?" in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3142–3152.

[10] 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," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 1477–1485.

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

[12] 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 *Proceedings of the European conference on computer vision (ECCV)*. Springer, 2022, pp. 1–18.

[13] Y. Wang, V. Guizilini, T. Zhang, Y. Wang, H. Zhao, , and J. M. Solomon, "Detr3d: 3d object detection from multi-view images via 3d-to-2d queries," in *The Conference on Robot Learning (CoRL)*, 2021.

[14] Y. Liu, T. Wang, X. Zhang, and J. Sun, "Petr: Position embedding transformation for multi-view 3d object detection," in *Proceedings of the European conference on computer vision (ECCV)*. Springer, 2022, pp. 531–548.

[15] S. Wang, X. Jiang, and Y. Li, "Focal-petr: Embracing foreground for efficient multi-camera 3d object detection," *IEEE Transactions on Intelligent Vehicles*, 2023.

[16] J. Huang and G. Huang, "Bevdet4d: Exploit temporal cues in multi-camera 3d object detection," *arXiv preprint arXiv:2203.17054*, 2022.

[17] 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," *arXiv preprint arXiv:2210.02443*, 2022.

[18] C. Han, J. Sun, Z. Ge, J. Yang, R. Dong, H. Zhou, W. Mao, Y. Peng, and X. Zhang, "Exploring recurrent long-term temporal fusion for multi-view 3d perception," *arXiv preprint arXiv:2303.05970*, 2023.

[19] Y. Liu, J. Yan, F. Jia, S. Li, A. Gao, T. Wang, and X. Zhang, "Petrv2: A unified framework for 3d perception from multi-camera images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3262–3272.

[20] X. Lin, T. Lin, Z. Pei, L. Huang, and Z. Su, "Sparse4d v2: Recurrent temporal fusion with sparse model," 2023.

[21] S. Wang, Y. Liu, T. Wang, Y. Li, and X. Zhang, "Exploring object-centric temporal modeling for efficient multi-view 3d object detection," *arXiv preprint arXiv:2303.11926*, 2023.

[22] Q. Zhu, C. Min, Z. Wei, Y. Chen, and G. Wang, "Deep learning for multi-view stereo via plane sweep: A survey," *arXiv preprint arXiv:2106.15328*, 2021.

[23] Z. Wei, Q. Zhu, C. Min, Y. Chen, and G. Wang, "Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6187–6196.

[24] G. Bae, I. Budvytis, and R. Cipolla, "Multi-view depth estimation by fusing single-view depth probability with multi-view geometry," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2842–2851.

[25] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[26] J. Deng, W. Dong, R. Socher, L. J. Li, and F. F. Li, "Imagenet: a large-scale hierarchical image database," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2009.

[27] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "Mvsnet: Depth inference for unstructured multi-view stereo," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 767–783.

[28] Y. Yao, Z. Luo, S. Li, T. Shen, T. Fang, and L. Quan, "Recurrent mvsnet for high-resolution multi-view stereo depth inference," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5525–5534.

[29] Y. Xue, J. Chen, W. Wan, Y. Huang, C. Yu, T. Li, and J. Bao, "Mvscrf: Learning multi-view stereo with conditional random fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4312–4321.

[30] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2495–2504.

[31] Z. Yu and S. Gao, "Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1949–1958.

[32] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3d object detection for autonomous driving," in *Proceedings*

of the IEEE/CVF conference on computer vision and pattern recognition, 2016, pp. 2147–2156.

[33] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *Proceedings of the European conference on computer vision (ECCV)*. Springer, 2020, pp. 194–210.

[34] J. Huang and G. Huang, "Bevpoolv2: A cutting-edge implementation of bevdet toward deployment," *arXiv preprint arXiv:2211.17111*, 2022.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[36] Y. Li, H. Bao, Z. Ge, J. Yang, J. Sun, and Z. Li, "Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 1486–1494.

[37] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017, pp. 2980–2988.

[38] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Computer Science*, 2014.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2016, pp. 770–778.