2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

3D Object Detection via Stereo Pyramid Transformers with Rich Semantic Feature Fusion

Rongqi Gu[1,2], Chu Yang[1,2], Yaohan Lu[2], Peigen Liu[1,2], Fei Wu[2], Guang Chen[1,3*]

***Abstract*— Camera-based 3D object detectors, prized for their broader applicability and cost-effectiveness compared to LiDAR sensors, still grapple with the inherently ill-posed nature of depth extraction from images. In this work, we present a novel approach that employs a transformer-based backbone and a fused geometry volume to bolster feature richness and elevate detection accuracy. Firstly, we propose the Stereo Pyramid Transformer backbone to extract features from stereo images, which can capture global information and establish cross-image semantic connections. Then, to tackle the challenge posed by small baseline binocular cameras, we propose to fuse stereo geometry volumes constructed by Stereo Plane Sweeping Volume (SPSV), Monocular Semantic Volume (MSV), and Lifted Volume (LV) to create finely detailed feature volumes. Through extensive experiments on both the KITTI and our datasets, our approach not only surpasses all existing transformer-based stereo 3D detection methods but also marks a significant milestone by achieving comparable performance with the leading CNN-based 3D detectors for the first time.**

I. INTRODUCTION

3D perception is a vital yet complex element in autonomous driving systems. While LiDAR-based techniques deliver accurate spatial data, their high expense and sparse resolution limit their commercial viability. In contrast, camera-based 3D object detection, with its superior resolution and affordability, has become a focal point of research [1]. However, the task of depth prediction from images presents significant challenges, impeding the construction of accurate 3D representations and leading to suboptimal outcomes in image-based 3D object detection.

To improve depth estimation accuracy from images, researchers have explored a variety of innovative strategies. Within the multi-view stereo (MVS) field, stereo matching has emerged as a key technique [2], [3]. These approaches aim to generate depth maps through the use of cost volumes, which effectively capture depth uncertainty. Furthermore, some binocular models directly derive depth maps [4], while others implicitly integrate depth information into geometric feature volumes, enriching the 3D representation with accurate depth cues. However, the implicit integration of depth information into geometric volumes suffers from a lack of explicit depth supervision, which compromises depth prediction accuracy and, consequently, the efficacy of 3D detection. The challenges are further magnified by small baseline binocular cameras, where even a minor disparity of one pixel can lead to significant depth inaccuracies. To overcome these obstacles, our approach introduces the use of geometry volumes derived from single images, accompanied by explicit depth supervision, which can guide and correct the stereo feature volumes.

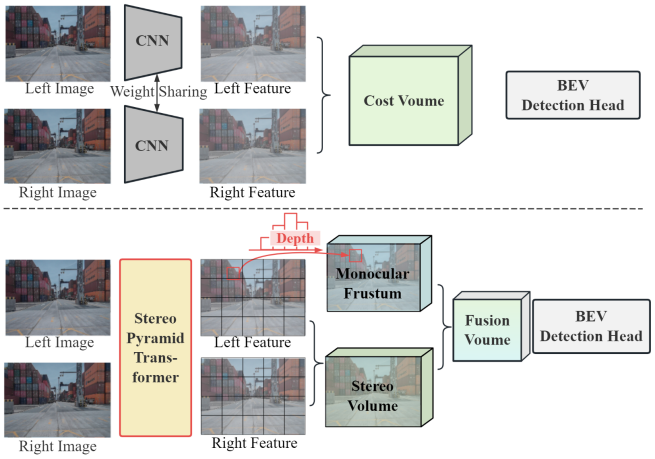

Fig. 1: Comparison of stereo detector frameworks. Existing stereo detectors adopt Siamese CNN backbone for separate feature extraction and cost volume construction, lacking comprehensive global information and cross-image features interaction. Our proposed framework introduces a stereo transformer, which can jointly model binocular features across global and multi-scale receptive fields.

Fig. 1 illustrates our innovative framework. Current stereo 3D detectors predominantly utilize convolutional neural networks (CNNs) within a Siamese network architecture to generate cost volumes from binocular images. However, the features extracted by CNNs often lack comprehensive global information, constrained by the networks' limited receptive fields. Furthermore, the structure of these Siamese networks, where the left and right image branches operate independently, precludes effective feature interaction between the two images. These limitations hinder the full exploitation of stereo imagery's potential for depth perception and 3D object detection. To this end, our proposed Stereo Pyramid Transformer (SPT) incorporates attention modules to obtain the global awareness capability and can jointly model binocular features across global and multi-scale receptive fields.

Our contribution encompasses three key aspects:

- Firstly, we propose the Stereo Pyramid Transformer (SPT), designed to forge cross-image relationships and achieve comprehensive global attention in binocular image pairs, outperforming traditional Siamese-style

*Guang Chen is the corresponding author, guangchen@tongji.edu.cn
Authors Affiliation: [1]Department of Automotive Engineering, Tongji University, Shanghai, China; [2]Shanghai Westwell Technology Co., Ltd, Shanghai, China; [3]Department of Computer Science and Technology, Tongji University, Shanghai, China.

transformer backbones.

- Secondly, we present a novel fused stereo volume approach for 3D geometric representation, specifically designed to enhance the detection accuracy of systems utilizing small baseline binocular cameras.
- Thirdly, we analyze our method's performance through comprehensive experimentation. The results unequivocally show that our model surpasses the majority of existing stereo-based detectors across both 3D and 2D metrics, thereby confirming its effectiveness.

II. Related work

A. Mono- and stereo-based depth estimation

Monocular-based depth estimation Eigen [5] firstly pioneered CNNs for depth prediction, outperforming traditional methods. With the development of neural networks in depth estimation, methods based on the monocular and multi-view stereo (MVS) merged [6]. For monocular methods, Monodepth [7] advances unsupervised monocular depth prediction using disparity and reconstruction, with Monodepth2 [8] incorporating depth estimation and pose networks for single-frame depth inference.

Stereo-based depth estimation DeepMVS [9] computes correlations between patches from reference and plane-sweep volumes to obtain fused matching features. Despite the rise of transformers, STTR [10] utilizes the global receptive field of the transformer to construct a correlation map between stereo pairs using regularized attention. However, the transformer's propensity to overfit on limited data poses a challenge, resulting in suboptimal performance for STTR on both the KITTI dataset and our proprietary dataset.

B. Image-based 3D object detection

Image-based 3D object detectors integrate monocular or multi-view depth estimation with object detection. Initially, methods rely on 2D results projected into 3D space, as exemplified by Stereo-RCNN [11].

An increasing trend is predicting objects directly in 3D space using monocular or multi-view stereo images. Monocular-based methods primarily use 2D or 3D features. 2D feature-based approaches, like LSS [12], estimate depth distribution and generate 3D frustum features, increasingly popular in recent studies [13]–[16]. Depth supervision from LiDAR improves accuracy in methods like BEVDepth [14].

3D feature-based techniques project image features into pre-generated 3D grid space [17], [18], e.g. DETR3D uses projected 3D object queries for feature collection.

In multi-view stereo images (MVS), PLUMENet [19] facilitates simultaneous object detection and occupancy prediction, while YoloStereo [20] enhances small object detection and reduces computational costs. DSGN and DSGN++ [21], [22] incorporate depth information into feature channels, aggregating 3D features from stereo pairs and monocular semantics. Different from DSGN++, we propose to leverage explicitly modeled monocular depth distribution for spatial feature construction.

C. Transformers for feature extraction

Transformers have revolutionized computer vision, with applications spanning various domains. They excel in capturing long-range dependencies among sequence elements compared to traditional CNNs. Transformer-based methods can be broadly categorized into three types [23]: (a) those employing a transformer backbone for feature extraction coupled with a detection head [24]–[26]; (b) those integrating a CNN backbone for feature extraction with a transformer-based decoder [27], [28]; and (c) end-to-end detection models [29]. In our study, we focus on leveraging a transformer backbone for feature extraction.

T2T [26] addresses the performance gap between traditional ViT and CNN backbones, particularly evident in smaller datasets, by combining neighboring tokens and aggregating spatial context. Swin Transformer [24] computes attention within confined yet changing windows, retrieving locality in the attention mechanism.

We opt for a transformer backbone over a CNN backbone in our work due to the significant size variances among objects in our dataset, for which the transformer's ability to capture long-range relationships is beneficial.

III. Method

A. Network overview

The framework of our detection model is depicted in Fig. 2. It integrates the Stereo Pyramid Transformer (SPT) and develops fused volumes in 3D space, which incorporates stereo plane sweeping volume (SPSV), monocular semantic volume (MSV), and lifted volume (LV) to enhance performance.

The Stereo Plane Transformer (SPT) processes stereo images to create the Stereo Plane Sweeping Volume (SPSV), Monocular Semantic Volume (MSV), and Lifted Volume (LV), which are transformed from frustum to 3D space for spatial consistency. These volumes are fused to form a 3D representation for accurate object detection, with deep supervision applied to the LV and SPSV to enhance depth estimation and 3D geometry construction.

B. Stereo Pyramid Transformer

Due to the significant differences in the shape and scale of targets in the dataset and the need for accurate depth estimation, CNN-based backbones, such as ResNet, struggle to extract rich features from binocular images. In this case, the transformer backbone shows a relative advantage since it assigns different patches with self-adaptive weights and maps features into multiple attention spaces, aiding in capturing global information. Inspired by [26], we propose the Stereo pyramid transformer module (SPT) to extract features from multi-view stereo images. Fig. 3 presents the architecture of our transformer backbone, the image patches are fed into the Stereo Token Attention Module (STA) to capture local structure information, where the adjacent tokens are aggregated to reduce computation cost. Then, the self-attention module with sinusoidal position encoding is applied to enrich global relationships among the resulting tokens.

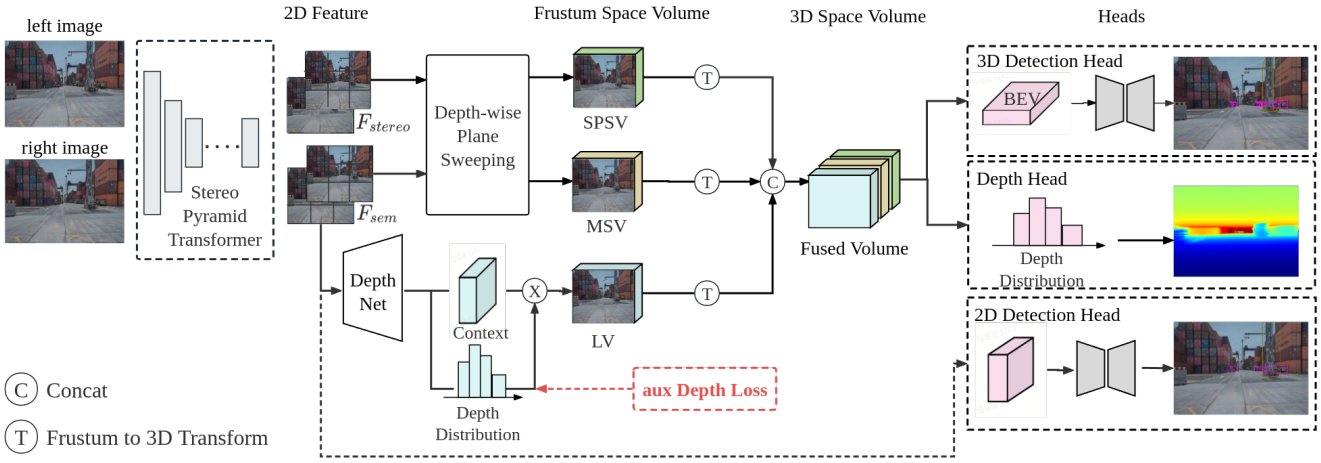

Fig. 2: The framework of our 3D detection model. Given input stereo images, the SPT backbone is used to extract semantic and stereo image features. These stereo image features are passed through plane sweeping to obtain SPSV. Mono depth module uses the left semantic feature as input and outputs LV in 3D space as described in III-C. MSV is constructed from the left semantic feature. Finally, these three volumes are fused by concatenating and sent to 3D and 2D object detection heads. Meanwhile, the depth head is used for supervising depth signals.

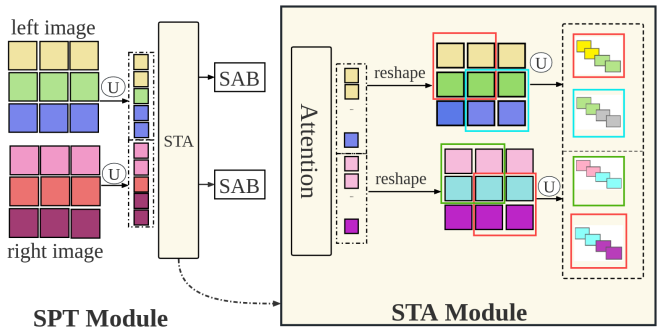

Fig. 3: Overview of Stereo Pyramid Transformer module (SAB denotes self attention block). The binocular images are split into patches and the STA module maps the concatenated patches as tokens for self-attention. Then the tokens are separated and transformed back to the image shape for the next STA stage.

Stereo Token Attention Module In binocular-based 3D detection, mainstream methods use a Siamese network with shared weights to extract features from the left and right images separately [21], [22], [30]–[32]. However, these approaches ignore the correlation between the binocular images. To this end, inspired by STTR [10] and T2T [26], we concatenate the left and right image patches together as input to the stereo token attention (STA) module to extract the global information from each image and cross relationships between images. The computation process can be described as:

$$T' = MLP(SA(T)) + SA(T) \tag{1}$$

where SA denotes the self-attention with one layer, MLP denotes multi-layer perceptron with GELU activate function. The tokens are normalized before being fed into the self-attention block. After that, the tokens T' can be reshaped into image shapes.

$$I_f = Reshape(T') \tag{2}$$

where $I_f \in \mathbb{R}^{h' \times w' \times c'}$. We utilize overlapping split restructured information [26] to aggregate the local information, aiming to make up for the inadequacy of the transformer. The concatenated tokens are divided into left and right sequences initially and reshaped back in image shape. For reshaped images, neighboring tokens will be split with overlapping into reduced tokens as input for the next STA module, such that the local information is aggregated.

C. Fused Volumes for 3D Geometry Representation

Capturing accurate and detailed 3D geometry from stereo camera-based image features is critical for 3d object detection. For 3D object detection networks, the primary parameters are dedicated to 2D image modeling, whereas the supervision signal is mostly defined in the 3D space. Due to such division, the 2D–3D conversion may stand as a bottleneck in the information flow during network training. Drawing on these concepts, our approach utilizes three 3D volumes, including Stereo Plane Sweep Volume (SPSV), Monocular Semantic Volume (MSV), and Lifted Volume (LV), to transform 2D image features into a joint 3D geometric representation, ensuring the accurate reconstruction of 3D geometric space.

Stereo Plane Sweep Volume Given the focal length f_u and the baseline b of a binocular camera, the disparity D of points located at depth plane with depth z can be calculated as:

$$D = \frac{f_u \times b}{z} \tag{3}$$

For every evenly spaced depth plane within the view frustum, the downsized disparities based on the feature map stride

can be computed. Concatenating the offset image features along channel dimension yields the plane-sweep cost volume $V_{cost} \in \mathbb{R}^{H \times W \times 2D \times C}$:

$$V_{cost}(u, v, d) = concat[F_L^{Stereo}(u, v), F_R^{Stereo}(u - \frac{D}{s}, v] \quad (4)$$

where s is the feature map stride of the backbone and F_L^{Stereo} and F_R^{Stereo} are stereo features produced by the stereo transformer image backbone. The raw cost volume is then processed with a stereo-matching network to form the Stereo Plane-Sweep Volume (SPSV).

$$V_{spsv} = ConvBN(ReLU(ConvBN(V_{cost}))) \quad (5)$$

We follow [22] to construct the Stereo Plane Sweep Volume in a depth-aware manner that correlates the feature channels with depth to enable richer information in image features. Specifically, image feature channels C_F are sliced by a sliding window of length $C_V (C_V \leq C_F)$ according to the corresponding depth plane of the volume. In this way, wider feature channels are enabled to carry richer depth-aware information about 3D space. We refer to [22] for detailed implementation. The SPSV provides geometry-guided 2D-3D transformation, ensuring the accurate recovery of 3D geometry space.

Monocular Semantic Volume We further adopt the depth-aware sliding window mechanism to construct monocular semantic volume (MSV) from semantic image features F_L^{Sem} following [22]. The MSV supplements the context information of the frustum space and proves to be beneficial. The semantic feature volume is constructed in the frustum space, thus will be transformed into 3D geometry space for consistency.

Lifted Volume In MSV, depth information is implicitly embedded into channels of the image feature. However, this coupling of depth and 2D features may lead to difficulty in feature learning. To this end, we seek explicit depth distribution modeling to facilitate more accurate 3D feature space reconstruction. Inspired by [12], we directly predict the depth distribution $P_d \in \mathbb{R}^{H \times W \times D}$ as probabilities of discrete depth bins from left semantic features with Monocular DepthNet. The predicted depth distribution is supervised by ground-truth depth provided by LiDAR points. The Lifted Volume (LV) $V_{LV} \in \mathbb{R}^{H \times W \times D \times C}$ are the outer production of the semantic feature $F_L^{Sem} \in \mathbb{R}^{H \times W \times C}$ and the depth distribution.

$$V_{LV} = (F_L^{Sem} \otimes P_d) \quad (6)$$

Different from SPSV and MSV, depth modeling and 2D visual feature learning are decoupled in the LV, in which the 3D feature space is constructed by weighting the complete image feature with explicitly modeled depth distribution along epipolar ray. The decoupling of the depth and feature learning eases the network training.

Fused Stereo Volume SPSV, MSV and LV share the same size in the camera frustum space. They are transformed into 3D space by sampling projected 3D grid points in normalized coordinates and are fused for more effective and robust 2D-3D transformation. For the fusion strategy, we concatenate these columns along the channel dimension and feed the fused volume into an hourglass network, which compresses and excites voxel features with 3D convolution and transposed convolutions. The resulting fused feature volume in 3D space can be used for downstream tasks including BEV detection and depth surface estimation.

IV. Experiment

A. Dataset

Our dataset contains a total of 8523 samples, divided into 6769 for training, 1754 for validation purposes.[1] It was collected by an autonomous unmanned vehicle operating within a port in Thailand, encompassing a diverse array of scenarios under various meteorological conditions. These conditions include bright daylight, nighttime with artificial lighting, cloudy days, dawn, and rainy periods, offering a comprehensive testing ground for the system. The binocular camera has a resolution of 900×500 with approximately 90° FOV and is mounted at the front of the vehicle at a height of about 1.5 meters, facing straight ahead, to support the 3D object detection in the forward direction. We manually labeled the 3D bounding boxes of objects in the LiDAR point clouds and subsequently transformed them into the camera coordinate system. Ground-truth depth maps are generated with LiDAR point clouds following [14].

Fig. 4: Scenarios of our dataset.

The dataset includes seven categories of object annotations, such as operational vehicles and port facilities like trucks, trailers, streetlights, and harbor cranes. Compared to conventional autopilot datasets, it features greater inter-class differences, realistic long-tailed distributions, and severe occlusion challenges, making it difficult for current 3D detection methods. Detailed statistics are in Table I, and typical scenarios are shown in Fig 4 [2]

We further evaluate our method on the KITTI 3D object detection dataset, which includes 3712 stereo image pairs and point clouds for training, 3769 for validation, and a test set of 7518 samples with limited access to submission online. The annotated objects are cars, pedestrians, and cyclists, each divided into three difficulties according to visibility.

[1]The dataset will be available upon request.

[2]For commercial protection, the texts in the image will be mosaiced.

TABLE I: Details of the dataset. The number of samples and the approximate size of each class are listed.

Class	Car	RTG	Truck head	Trailer	Street lamp	Lock station	QC
3D bounding boxes	7546	6378	12800	16444	5469	4712	4703
Dimensions(LWH)	5.6, 2, 2.5	14, 1.9, 3	4.6, 3, 3.5	13.5, 3, 3.3	5.6, 2.5, 3	3.4, 1.7, 2.5	26.7, 3.2, 8

B. Training and Inference

Given the scarcity of supervision signals in 3D object detection, we leverage 2D tasks during training to bolster 2D feature modeling. Specifically, we concurrently address three tasks—3D target detection, 2D target detection, and image depth estimation simultaneously during training. Depth supervision is imposed on the point-wise depth distribution from the monocular depth net as well as the final depth map predicted from the depth head using fused 3D volume. Smooth L1 Loss and RDIoU [33] loss are employed for 3D target box regression, while cross-entropy loss is utilized for target orientation classification. Focal Loss is applied for the classification of unbalanced categories. Depth estimation is supervised with the cross-entropy loss with linear label smoothing. 2D detection involves conventional GIOU loss and focal loss. Common data augmentations including random flip, crop, rotation, and scaling are adopted in training.

C. Experimental Setup

Evaluation Metric. We evaluate our method using a simplified KITTI metric since the complexity of the KITTI metric could be rather confusing when evaluating more categories in our dataset. KITTI defines three difficulty levels based on front-view images and uses different IOU thresholds for different categories for recall-precision plotting. The precision scores over 40 recall rates are averaged to get $AP|_{R40}$. We only focus on Moderate-difficulty targets with an IoU threshold of 0.5 for all categories. We further report the mean average precision (mAP) similar to NuScenes [34]:

$$mAP = \frac{1}{C}\sum_{c \in C} AP^c|_{R40} \tag{7}$$

Baseline Methods. The proposed framework adopts the main structure of DSGN++ [22]. To validate the effectiveness of our method, we further include several other strong baselines for comparison, including DSGN [21], YoloStereo3D [20] and LIGA-Stereo [35]. We adapt the baseline methods to our dataset following their default configuration for a fair comparison. It is worth noticing that LIGA-Stereo leverages LiDAR point clouds input to perform feature-wise imitation learning to improve performance, which is not involved in training our model.

Implementation Details. The input image size of our model is $(W_I = 928, H_I = 480)$. The whole transformer backbone contains 2 parts, the Stereo Pyramid Transformer module and the classic self-attention transformer backbone. Here we set the number of SPT modules to 2, which means there are 3 overlapped-split operations and 2 attention blocks. The patch size for 3 overlapped-split operations is $[7, 3, 3]$, the overlapping stride is $[4, 2, 2]$. The input image size is down-sampled from $[480, 928]$ to $[30, 58]$ after 2 SPT modules. In order to enrich the global relationship among tokens, we set 7 layers self-attention blocks for the left and right features respectively, which are the same as the Token-to-Token ViT [26], allowing the pre-trained model weights to converge faster.

Table I highlights imbalanced target bounding box distribution in our dataset. To deal with the influence of the long-tailed distribution, we augment small-scale class by copying from other frames and transforming to the current frame. Common copy-paste methods randomly paste objects onto 2D images risk damaging depth information. To tackle this, we utilize the relative location in LiDAR coordinate from other frames to project a 2D bounding box onto the image. Additionally, to avoid destroying the original scene and objects, we filter the augmented objects by the convex placeable area mined from point clouds. To this end, the height of augmented boxes is adjusted by the road plane information fitted from the placeable region using the RANSAC algorithm.

Our model presents a voxelized 3D space in the range $(X \in [2, 59.6], Y \in [-30.4, 30.4], Z \in [-3, 1]$ (axis aligned with KITTI LiDAR coordinate system except for centering at the left stereo camera) with voxel size $0.2m \times 0.2m \times 0.2m$. The constructed volume was downsampled four times in width, height, and depth dimensions to reduce computational costs. The proposed model is trained with NVIDIA RTX 3090 GPUs for 60 epochs. The adopted optimizer is AdamW with one warm-up epoch. The initial learning rate is set to $1e-3$ with a weight decay of $1e-4$ and divided by 10 at the last 10 epochs.

D. Quantitative Results

We report experimental results with comparison on our dataset, as shown in Tab. II and III. It can be observed that our method outperforms all other state-of-the-art stereo-based approaches over all categories. Compared to the baseline method DSGN++, the improvement is notably substantial ($mAP_{3D} + 8.82, mAP_{BEV} + 7.75, mAP_{2D} + 7.78, mAP_{AOS} + 7.88$), validating the effectiveness of our approach.

Regarding detection accuracy across various categories, our method exhibits the highest accuracy in all 7 categories, and the discrepancy in detection performance between our method and the baseline approach is especially pronounced in the QC and Lock station categories (10.5% and 24.83% over DSGN++), which are often located far away or heavily occluded. We attribute the strengthened capability of localizing distant objects primarily to the enhanced frontward feature by LV, as evidenced by similar enhancements in LIGA-Stereo. Our model performs better for larger objects like RTG and QC (65.78% and 64.49% in mAP_{3D}) due to the global receptive field brought by the proposed SPT backbone, while the accuracy for smaller objects, such as

Method	Truck head		RTG		Tray		QC		Car		Lock station		Street lamp	
	AP_{3D}	AP_{BEV}	AP_{3D}	AP_{BEV}	AP_{3D}	AP_{BEV}	AP_{3D}	AP_{BEV}	AP_{3D}	AP_{BEV}	AP_{3D}	AP_{BEV}	AP_{3D}	AP_{BEV}
YoloStereo3D [20]	20.64	23.45	46.55	47.49	37.53	42.16	45.17	50.59	16.37	20.69	10.43	13.15	30.53	34.07
DSGN [21]	10.80	12.34	42.83	43.09	37.71	42.50	37.55	37.74	1.41	2.18	6.36	8.85	45.69	47.33
LIGA-Stereo† [35]	36.62	41.20	56.99	58.36	57.06	63.90	58.02	59.51	6.27	9.06	29.50	33.75	40.08	41.15
DSGN++ [22]	42.59	46.23	64.00	64.68	59.96	65.62	53.99	55.56	13.81	18.87	13.39	21.33	55.68	56.45
Ours	**45.19**	**48.73**	**65.78**	**65.95**	**65.20**	**68.83**	**64.49**	**65.74**	**25.87**	**30.07**	**38.22**	**42.79**	**60.41**	**60.83**

TABLE II: Detailed results on our dataset. †indicates that the model utilizes LiDAR feature imitation.

Method	mAP_{3D}	mAP_{BEV}	mAP_{2D}	mAP_{aos}
YoloStereo3D [20]	29.61	33.09	54.81	50.23
DSGN [21]	26.05	27.72	21.77	20.13
LIGA-Stereo [35]	40.65	43.85	67.21	62.01
DSGN++ [22]	43.35	46.96	66.06	61.45
Ours	**52.17**	**54.71**	**73.84**	**69.33**

TABLE III: Overall results on our dataset. Mean Average Precision (IOU=0.5) for all 7 classes.

Lock station (38.22% in mAP_{3D}), are comparatively lower.

To further validate our approach, we also conducted experiments on the KITTI benchmark, and the results are summarized in Tab. IV. TS3D [39] is the first public transformer-based stereo 3D detection approach, which adopts an encoder-decoder architecture with disparity embedded into stereo features. Our model demonstrates significant superiority over TS3D, achieving a 10% higher AP_{3D}, attributed to the construction of 3D spatial representation in a geometry-aware manner. Compared to CNN-based methods, our model performs comparably with the state-of-the-art methods on both validation and test sets.

Inference Time The inference time of our approach is measured on an NVIDIA RTX 3090 GPU. The forward propagation of the network takes 0.29s on average and the fused 2D-3D transform takes 0.22s.

E. Qualitative Results

To demonstrate the effectiveness of our proposed model more intuitively, we present a series of qualitative results in Fig. 5 and Fig. 6. These examples span an array of scenario conditions, including fluctuating illumination, obstructions, distant objects and polluted camera sight.

In Fig. 5, the first row depicts the predicted 3D bounding boxes, with distinct colors denoting different categories. The second row shows the corresponding depth estimations. It can be observed that our model demonstrates a remarkable capability to produce highly accurate 3D bounding boxes for a range of targets, even in the presence of significant challenges such as severe occlusions, rainy conditions, and polluted sight. Besides, the predicted depth maps exhibit sufficient precision, enhancing the performance of 3D object detection in dynamic environments.

In Fig. 6, we present a comparison of the qualitative results between our proposed method and the baseline approach. It is evident that our model delivers more consistent detection performance under extreme lighting conditions. Additionally, it significantly reduces the missed detections of occluded and distant objects.

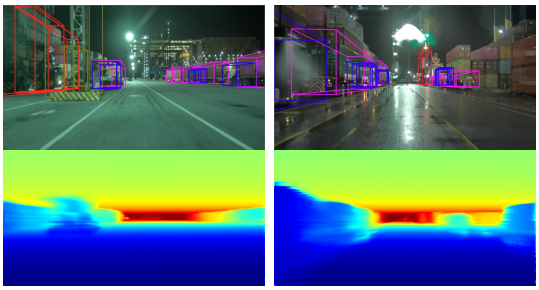

Fig. 5: Detection results of our proposed method. The first row depicts 3d bounding boxes. The second row depicts a colored depth prediction map.

F. Ablation Studies

In this section, we validate the effectiveness of our model design by ablating proposed modules. The experiment results are illustrated in Table V, VI and VII. We introduce median depth error over the positions overlaid with ground truth to further analyze the performance of the methods.

Effect of 3D Volume Fusion We ablate the feature volumes separately and compare the SPT to the CNN backbone to assess their effectiveness. As illustrated in Tab. V, the removal of LV results in a significant decline in performance, with a decrease of (-4.93% in mAP_{3D} and -4.71% in mAP_{BEV}) and greater depth error. This accentuates the advantage of introducing depth supervision throughout the 3D feature volume's construction over mere postposition, substantially boosting the accuracy and performance of the ultimate 3D detection. The adoption of MSV enhances the semantic features in fused representation, bringing $+0.81\%$ improvement in mAP_{3D}.

Stereo Pyramid Transformer Designs The effectiveness of the SPT module is discussed in Table VI. We compare the proposed STA module with the Siamese-style variant where left and right tokens are separately processed. Specifically, the STA-S module does not perform connection operations, and attention and overlapping split operations are performed on each image, thus preventing the cross-relationship of stereo features. The experimental results show that the cross-connection of stereo tokens greatly boosts the model performance, increasing the detection accuracy by $+18.17\%mAP_{3D}, +19.41\%mAP_{BEV}$ and significantly reducing the depth error from 0.25 to 0.1567.

The proposed transformer-based backbone is further compared with two strong transformer backbones, i.e. Swin [24] and TransXNet [40]. All backbones are trained with pre-trained weights for a fair comparison. As demonstrated in Tab. VII, the proposed SPT shows superior performance over

Sensor	Method	Type	*val* set AP_{3D}			*val* set AP_{BEV}			*test* set AP_{3D}			*test* set AP_{BEV}		
			Easy	Mod	Hard	Easy	Mod	Hard	Easy	Mod	Hard	Easy	Mod	Hard
LiDAR	SECOND [36]	-	87.43	76.48	69.10	89.96	87.07	79.66	83.34	72.55	65.82	89.39	83.77	78.59
	PointPillars [37]	-	-	-	-	-	-	-	82.58	74.31	68.99	90.07	86.56	82.81
	PV-RCNN [38]	-	92.57	84.43	82.69	95.76	91.11	88.93	90.25	81.43	76.82	94.98	90.65	86.14
Stereo	YoLoStereo3D [20]	CNN	72.06	46.58	35.53	-	-	-	65.68	41.25	30.42	76.10	50.28	36.86
	DSGN [21]	CNN	72.31	54.27	47.71	83.24	63.91	57.83	74.52	54.22	46.36	83.32	66.24	57.65
	LIGA [35]	CNN	84.92	67.06	63.80	89.35	77.26	69.05	81.39	64.66	57.22	88.15	76.78	67.40
	DSGN++ [22]	CNN	-	69.12	-	-	78.93	-	83.21	67.37	59.91	88.55	78.94	69.74
	TS3D [39]	T	70.90	46.76	35.94	-	-	-	64.61	41.29	30.68	-	-	-
	Ours	T	85.14	68.18	62.72	91.74	77.36	72.02	82.03	64.60	56.73	88.39	75.79	66.51

TABLE IV: Car detection results on KITTI *val* set and *test* set. Type T denotes Transformer-based.

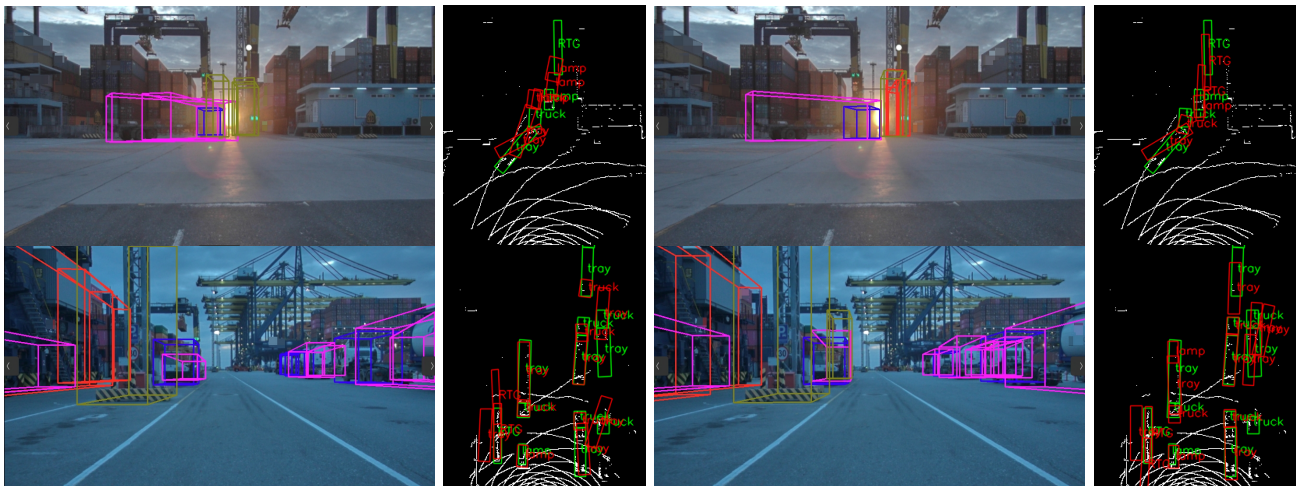

Fig. 6: Comparison of detection results. The left part is the baseline results. The right part is the results of our method.

ID	Method	SPT	MSV	LV	$mAP_{3D}\uparrow$	$mAP_{BEV}\uparrow$	$mAP_{2D}\uparrow$	$mAP_{aos}\uparrow$	depth error↓
1	Ours	✓	✓	✓	**52.17**	**54.71**	**73.84**	69.33	**0.1472**
2	Ours w/o LV	✓	✓	✗	47.24	50.00	70.27	66.35	0.1648
3	Ours w/o MSV	✓	✗	✓	51.36	54.40	73.82	**69.95**	0.1517
4	CNN backbone (ResNet-34)	✗	✓	✓	50.80	54.66	72.95	68.84	0.1582
5	DSGN++ (baseline)	✗	✓	✗	43.35	46.96	66.06	61.45	0.17

TABLE V: Ablation studies on network modules. SPT denotes applying the Stereo Pyramid Transformer for feature extraction or the original Siamese CNN Backbone.

Backbone	AP_{3D}	AP_{BEV}	AP_{2D}	AP_{aos}	depth error↓
STA-S	27.39	28.87	40.60	36.28	0.25
STA	**45.56**	**48.29**	**69.16**	**63.76**	**0.1567**

TABLE VI: Ablation studies on feature extraction backbone on *validation* set. STA denotes Stereo Pyramid Transformer. STA-S denotes the Siamese variant of STA that operates on the binocular images separately.

Backbone	AP_{3D}			AP_{BEV}		
	Easy	Mod	Hard	Easy	Mod	Hard
Swin [24]	79.18	61.11	54.48	88.09	70.29	64.72
TransXNet [40]	81.46	62.33	56.84	88.22	71.09	65.78
SPT	**85.14**	**68.18**	**62.72**	**91.74**	**77.36**	**72.02**

TABLE VII: Alabtion Studies on transformer backbone on KITTI val set.

Swin and TransXNet on the KITTI dataset and led the Swin-transformer by 5.85% / 6.27% in mAP_{3D} / mAP_{BEV}.

V. CONCLUSIONS

In this paper, we propose a novel transformer-based framework for 3D object detection from stereo images. We propose the Stereo Pyramid Transformer, which can promote early interaction between stereo pairs to enhance 2D modeling. We also introduce the fusion of stereo- and mono-based feature volumes, and explicit depth supervision to significantly improve the depth estimation accuracy to aid object detection. Extensive experiments demonstrate that our approach achieves superior accuracy and robustness.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (No. 62372329), in part by Shanghai Scientific Innovation Foundation (No.23DZ1203400), in part by Shanghai Rising Star Program (No.21QC1400900), in part by Tongji-Qomolo Autonomous Driving Commercial Vehicle Joint Lab Project, and in part by Xiaomi Young Talents Program.

REFERENCES

[1] Y. Cai, T. Zhang, H. Wang, Y. Li, Q. Liu, and X. Chen, "3d vehicle detection based on lidar and camera fusion," *Automotive Innovation*, vol. 2, pp. 276–283, 2019.

[2] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5410–5418, 2018.

[3] 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*, pp. 2495–2504, 2020.

[4] A. D. Pon, J. Ku, C. Li, and S. L. Waslander, "Object-centric stereo matching for 3d object detection," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8383–8389, IEEE, 2020.

[5] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in neural information processing systems*, vol. 27, 2014.

[6] G. Li, X. Chi, and X. Qu, "Depth estimation based on monocular camera sensors in autonomous vehicles: A self-supervised learning approach," *Automotive Innovation*, vol. 6, no. 2, pp. 268–280, 2023.

[7] P.-Y. Chen, A. H. Liu, Y.-C. Liu, and Y.-C. F. Wang, "Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 2624–2632, 2019.

[8] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3828–3838, 2019.

[9] P.-H. Huang, K. Matzen, J. Kopf, N. Ahuja, and J.-B. Huang, "Deepmvs: Learning multi-view stereopsis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2821–2830, 2018.

[10] Z. Li, X. Liu, N. Drenkow, A. Ding, F. X. Creighton, R. H. Taylor, and M. Unberath, "Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6197–6206, 2021.

[11] P. Li, X. Chen, and S. Shen, "Stereo r-cnn based 3d object detection for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7644–7652, 2019.

[12] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pp. 194–210, Springer, 2020.

[13] C. Reading, A. Harakeh, J. Chae, and S. L. Waslander, "Categorical depth distribution network for monocular 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8555–8564, 2021.

[14] 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, pp. 1477–1485, 2023.

[15] Z. Li, Z. Yu, W. Wang, A. Anandkumar, T. Lu, and J. M. Alvarez, "Fb-bev: Bev representation from forward-backward view transformations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6919–6928, 2023.

[16] 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.

[17] 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 *European conference on computer vision*, pp. 1–18, Springer, 2022.

[18] 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 *Conference on Robot Learning*, pp. 180–191, PMLR, 2022.

[19] Y. Wang, B. Yang, R. Hu, M. Liang, and R. Urtasun, "Plumenet: Efficient 3d object detection from stereo images," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3383–3390, IEEE, 2021.

[20] Y. Liu, L. Wang, and M. Liu, "Yolostereo3d: A step back to 2d for efficient stereo 3d detection," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13018–13024, IEEE, 2021.

[21] Y. Chen, S. Liu, X. Shen, and J. Jia, "Dsgn: Deep stereo geometry network for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12536–12545, 2020.

[22] Y. Chen, S. Huang, S. Liu, B. Yu, and J. Jia, "Dsgn++: Exploiting visual-spatial relation for stereo-based 3d detectors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4416–4429, 2022.

[23] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM computing surveys (CSUR)*, vol. 54, no. 10s, pp. 1–41, 2022.

[24] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

[25] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 568–578, 2021.

[26] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 558–567, 2021.

[27] W. Guo, Z. Li, Y. Yang, Z. Wang, R. H. Taylor, M. Unberath, A. Yuille, and Y. Li, "Context-enhanced stereo transformer," in *European Conference on Computer Vision*, pp. 263–279, Springer, 2022.

[28] B. Li, Y. Sun, X. Jin, W. Zeng, Z. Zhu, X. Wang, Y. Zhang, J. Okae, H. Xiao, and D. Du, "Stereoscene: Bev-assisted stereo matching empowers 3d semantic scene completion," *arXiv preprint arXiv:2303.13959*, 2023.

[29] Y. Fang, B. Liao, X. Wang, J. Fang, J. Qi, R. Wu, J. Niu, and W. Liu, "You only look at one sequence: Rethinking transformer in vision through object detection," *Advances in Neural Information Processing Systems*, vol. 34, pp. 26183–26197, 2021.

[30] 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)*, pp. 767–783, 2018.

[31] Z. Qin, J. Wang, and Y. Lu, "Triangulation learning network: from monocular to stereo 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7615–7623, 2019.

[32] X. Cheng, Y. Zhong, M. Harandi, T. Drummond, Z. Wang, and Z. Ge, "Deep laparoscopic stereo matching with transformers," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 464–474, Springer, 2022.

[33] H. Sheng, S. Cai, N. Zhao, B. Deng, J. Huang, X.-S. Hua, M.-J. Zhao, and G. H. Lee, "Rethinking iou-based optimization for single-stage 3d object detection," in *European Conference on Computer Vision*, pp. 544–561, Springer, 2022.

[34] 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*, pp. 11621–11631, 2020.

[35] X. Guo, S. Shi, X. Wang, and H. Li, "Liga-stereo: Learning lidar geometry aware representations for stereo-based 3d detector," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3153–3163, 2021.

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

[37] 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*, pp. 12697–12705, 2019.

[38] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "Pv-rcnn: Point-voxel feature set abstraction for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10529–10538, 2020.

[39] H. Sun, Y. Pang, J. Cao, J. Xie, and X. Li, "Transformer-based stereo-aware 3d object detection from binocular images," *arXiv preprint arXiv:2304.11906*, 2023.

[40] M. Lou, H.-Y. Zhou, S. Yang, and Y. Yu, "Transxnet: Learning both global and local dynamics with a dual dynamic token mixer for visual recognition," *arXiv preprint arXiv:2310.19380*, 2023.