MonoDSSMs: Efficient Monocular 3D Object Detection with Depth-Aware State Space Models

Kiet Dang Vu[1,2][0009−0004−0964−8134], Trung Thai Tran[1,2][0009−0002−1422−9685], and Duc Dung Nguyen[1,2⋆][0000−0001−7321−7401]

[1] AITech Lab., Ho Chi Minh City University of Technology (HCMUT)
[2] Vietnam National University Ho Chi Minh City (VNUHCM)
{kiet.dangvutuan0712, thai.tran241002, nddung}@hcmut.edu.vn

Abstract. Monocular 3D object detection has been an important part of autonomous driving support systems. In recent years, we have seen enormous improvement in both detection quality and runtime performance. This work presents MonoDSSM, the first to utilize the Mamba architecture to push the performance further while maintaining the detection quality. In short, our contributions are: (1) introduce Mamba-based encoder-decoder architecture to extract 3D features, and (2) propose a novel Cross-Mamba module to fuse the depth-aware features and context-aware features using the State-Space-Models (SSMs). In addition, we employ the multi-scale feature prediction strategy to enhance the predicted depth map quality. Our experiments demonstrate that the proposed architecture yields competitive performance on the KITTI dataset while significantly improving the model's effectiveness in both model size and computational cost. Our MonoDSSM achieves a comparable detection quality to the baseline, with 2.2x fewer parameters and a 1.28x faster computation time.

Keywords: 3D Object Detection · Mamba · State Space Models

1 Introduction

The ability to precisely locate and identify objects in 3D space underpins a revolution in various fields. From navigating self-driving cars in complex environments to guiding robots in intricate tasks, 3D object detection serves as the cornerstone, empowering these applications with an accurate understanding of their surroundings. Past research leveraging LiDAR [12,18,24,43,44,57,63] and multi-camera [20,27,32,54,56] setups have yielded excellent results due to their detailed depth of information. However, these approaches still face some limitations. They depend on multiple sensors, which makes them susceptible to failure with inappropriate settings and unsuitable for budget-conscious deployments. Therefore, monocular 3D detection algorithms stand out as a promising alternative. They offer a robust and practical solution for scenarios with limited resources by requiring only a single camera.

⋆ Corresponding author.

Despite impressive progress, recent monocular 3D object detection methods [2,5,8,23,26,31,35,36] remain limited by the absence of depth cues. Several works have concentrated on producing estimated depth maps and using them to aid the learning detection process [3,11,21,52,60] to overcome this problem. Also, recent work MonoDTR [21] focuses on improving the context features extracted by the model's backbone by combining them with depth-aware information based on Transformer architecture. While these methods improved object localization through estimated depth, they still faced two challenges. Firstly, the overall optimization process of 3D object detection is sensitive to depth information, training on the guidance of inaccurate depth maps can lead to sub-optimal 3D detection performance. Secondly, with Transformer-based architecture, handling depth and context information effectively can produce large computational overhead.

To address these challenges, we introduce MonoDSSMs, an efficient monocular 3D object detection framework with a novel Mamba-based architecture. Inspired by the recent success of Mamba [10,16], a novel State Space Models (SSMs) [9,10,16,17,38,47] in capturing long-range dependencies, we introduce an encoder-decoder architecture based on Mamba (Fig. 2a). We propose BiMamba2 (Fig. 2b), a bidirectional-scan Mamba model to capture long-range visual features. By traversing in two routes, we can obtain global information with only much smaller features (DLA-34) than those extracted from the larger backbone (DLA-102) when using Transformer. We present CrossMamba (Fig. 2b) replacing the Attention mechanism in Transformer for the decoder, enabling MonoDSSMs to effectively fuse both depth and contextual features while maintaining computational efficiency. Also, to improve object depth estimation, we employ a multi-scale feature enhancement strategy that generates more precise depth cues (Fig. 4). We summarize our contributions as follows:

- We propose MonoDSSMs, an efficient monocular 3D object detection framework that utilizes Mamba-based encoder-decoder architecture. To the best of our knowledge, we are the first to leverage Mamba architecture in supporting the monocular 3D object detection task. With a simple scan strategy, global context can be captured with much smaller features, which can boost our computational efficiency while maintaining a competitive detection performance.
- We introduce CrossMamba, a novel architecture for feature fusion based on Mamba to integrate context and depth-aware features efficiently. Our module serves as an alternative to the Cross-Attention mechanism, which fuses the queried depth information to obtain finer features.
- We also take advantage of a multi-scale features depth prediction strategy to achieve a more precise depth map which can improve detection performance by quality depth hints.

2 Related Work

Monocular 3D Object Detection. Monocular 3D object detection faces a significant challenge: the lack of depth information in a single image leads to inaccurate object localization. Recent research has focused on improving models' ability to predict object depth. There are two main approaches to depth estimation: direct regression [62] and geometry-based depth derived from the pinhole camera model [4]. Building on these methods and the success of incorporating uncertainty with depth estimation [5,36,45], several studies have explored predicting both types of depth simultaneously and utilizing uncertainty fusion to achieve a more accurate final depth value [28,61,65]. Additionally, leveraging ground plane information has shown promising results [3,39,55], especially in addressing the ill-pose depth estimation from monocular images.

Depth-assisted Monocular 3D Object Detection. To further improve the performance, many approaches propose using depth information to aid 3D object detection [3,11,21,52,60]. Early approaches focused on specialized convolution methods to combine visual features with depth maps. For instance, Ding *et al*. [11] introduced a novel depth-guided filtering module to leverage the benefits of predicted depth maps, while Bui *et al*. [3] employed pixel-adaptive convolution [48] to seamlessly integrate depth information and guide the learning process across all output channels. With the rise of transformers in various tasks, recent works like MonoDTR [21] and MonoDETR [60] have achieved significant improvements by utilizing discrete depth maps and encoder-decoder transformer architectures.

Transformer. Originally introduced for sequential modeling in natural language processing (NLP), the Transformer architecture [51] has revolutionized the field with its impressive performance. The key to Transformer's success is its self-attention mechanism, which allows the model to capture long-range dependencies within the data. This capability has proven highly effective not only in NLP but also in recent visual recognition tasks [13,34,50]. The application of Transformers has even extended to monocular 3D object detection, achieving promising results [21,60]. This success paves the way for exciting new directions in 3D object detection research.

State Space Models. While Transformers have achieved impressive results across various domains and tasks, their self-attention mechanism suffers from quadratic complexity. This becomes a major bottleneck when handling long sequences like lengthy sentences or high-resolution images. To address this challenge, State Space Models (SSMs) have recently emerged as a promising alternative for managing long-range dependencies [9,10,16,17,38,47]. Similar to Transformers, which initially thrived in language tasks, SSMs or Mamba [10,16] are demonstrating potential in computer vision as well. However, the original Mamba block is designed for the 1-D sequence, which is not suitable for vision tasks requiring spatial-aware understanding. To address this, [19,22,25,29,33,64] proposed several scan directions to capture the spatial relationships between pixels within an image, essential for accurate image representation. In this work, we propose a novel architecture for monocular 3D object detection that employs

Mamba's ability to capture long-range dependency features and improve model efficiency.

3 MonoDSSMs

3.1 Overall Architecture

Fig. 1 illustrates the overview of our proposed framework. We use a small architecture, DLA-34 [58] as our backbone to extract visual features. Following MonoDTR [21], we then adopted two branches to parallel extract depth-aware and context-aware features. In the depth-aware branch, the MSDFE module (Sec. 3.4) is presented to learn depth-aware features through auxiliary discretization depth supervision. On the other hand, several convolution layers are applied in the context-aware branch to extract the needed features. Then, to integrate these two kinds of features, we proposed a novel Mamba-based encoder-decoder architecture (Sec. 3.3) and follow [21] to utilize depth-positional hint to the context-aware feature a depth positional encoding. Finally, we adopt a single-stage detector with prior-based 2D-3D anchor boxes [30, 41] and loss from [21] for 3D object detection.

3.2 Efficient Feature Extractor

Backbone. Previous works [21] adopted DLA-102 [58], a quite large model with 33M parameters as the vision backbone to extract features. This choice is

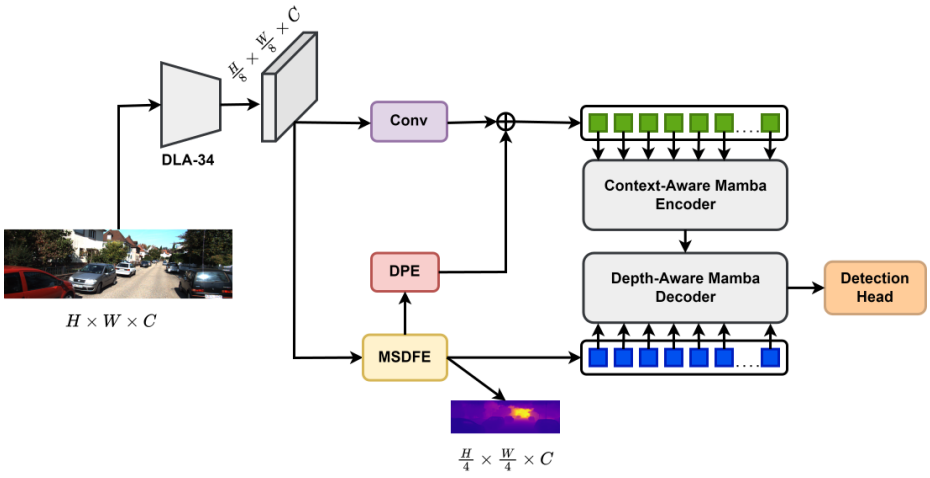


Fig. 1: The overall architecture of our proposed MonoDSSMs. After feeding the input image to the vision backbone, the MSDFE module is used to enhance the depth-aware features, while several convolutions are applied to extract the context-aware features. The Mamba encoder-decoder architecture is then employed to fuse these features. Finally, anchor-based detection is applied to obtain the results.

the main reason for forming inefficient models. To enhance the efficiency of the model, we found that DLA-34 [58] can be an ideal vision backbone for a fast and lightweight model while maintaining precision.

Feature Size. Given the input RGB image with resolution $H \times W$, both DLA-34 and DLA-102 output a feature map with resolution $\frac{H}{8} \times \frac{W}{8}$. The main difference is that DLA-34 produces a feature map with 128 channels while DLA-102 is 256. They then use this feature map to extract the depth/context-aware features and then are fused by an encoder-decoder architecture. Such a large feature map can result in computation overheads.

We found that integrating MonoDTR [21] with DLA-34 [58] as vision backbone and a small feature size $C = 128$ can build up a lightweight model with fewer parameters and faster inference time, but a drop in performance is inevitable as described in Tab. 2. However, we noticed that this slight decrease can be compensated by a model that can extract richer information with a similar number of features compared to Transformers.

3.3 Depth-Aware Mamba

Preliminaries. Structure State Space Models (SSMs) is a sequence model that can map a one-dimensional sequence $\mathbf{x} \in \mathbb{R}^L$ to $\mathbf{y} \in \mathbb{R}^L$ through a hidden state $\mathbf{h} \in \mathbb{R}^{L \times N}$ so that:

$$\begin{aligned} \mathbf{h}'(t) &= \mathbf{A}\mathbf{h}(t) + \mathbf{B}\mathbf{x}(t), \\ \mathbf{y}(t) &= \mathbf{C}\mathbf{h}(t) \end{aligned} \tag{1}$$

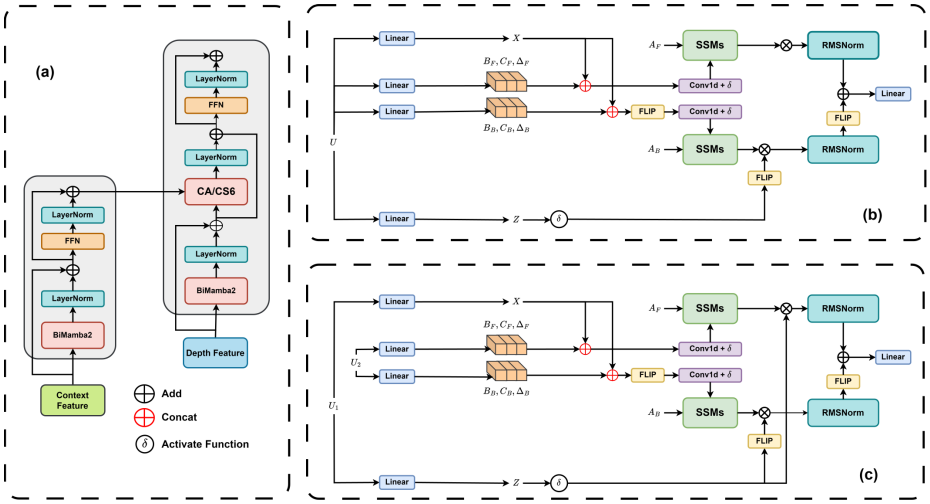


Fig. 2: The proposed Depth Aware Mamba. The core of our proposed MonoDSSMs architecture is the integration of context and depth features. It employs (a) a Mamba-based encoder-decoder architecture, (b) the proposed BiMamba2 block, which extracts visual context using two different scan routes, and (c) the CrossBiMamba2 module, which fuses the features based on Bidirectional Mamba.

where $\mathbf{A} \in \mathbb{R}^{N\times N}$, $\mathbf{B} \in \mathbb{R}^{N\times 1}$, $\mathbf{C} \in \mathbb{R}^{1\times N}$ are parameters of models.
Usually, when working on deep learning, the continuous system in Eq. (1) is discretized by time scale parameter $\mathbf{\Delta} \in \mathbb{R}$ to converting continuous variable $\mathbf{A}$, $\mathbf{B}$ to their discrete counterparts $\overline{\mathbf{A}}$, $\overline{\mathbf{B}}$ by ZOH methods:

$$\begin{aligned} \overline{\mathbf{A}} &= \exp\left(\mathbf{\Delta A}\right) \\ \overline{\mathbf{B}} &= \left(\mathbf{\Delta A}\right)^{-1}\left(\overline{\mathbf{A}} - \mathbf{I}\right) \cdot \mathbf{\Delta B} \end{aligned} \tag{2}$$

This leads to a linear-time invariant system formulation as follows:

$$\begin{aligned} \mathbf{h}_t &= \overline{\mathbf{A}}\mathbf{h}_{t-1} + \overline{\mathbf{B}}\mathbf{x}_t \\ \mathbf{y}_t &= \mathbf{C}\mathbf{h}_t \end{aligned} \tag{3}$$

To enhance computation efficiency, the SSMs model usually uses a global convolution operator that can take advantage of parallel computation:

$$\mathbf{y} = \mathbf{x} \circledast \overline{\mathbf{K}} \tag{4}$$

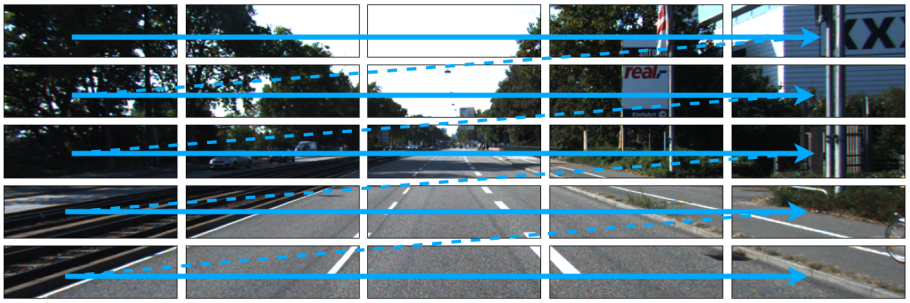
with

$$\overline{\mathbf{K}} = \left(\mathbf{C}\overline{\mathbf{B}}, \mathbf{C}\overline{\mathbf{A}\mathbf{B}}, \cdots, \mathbf{C}\overline{\mathbf{A}}^{L-1}\overline{\mathbf{B}}\right) \tag{5}$$

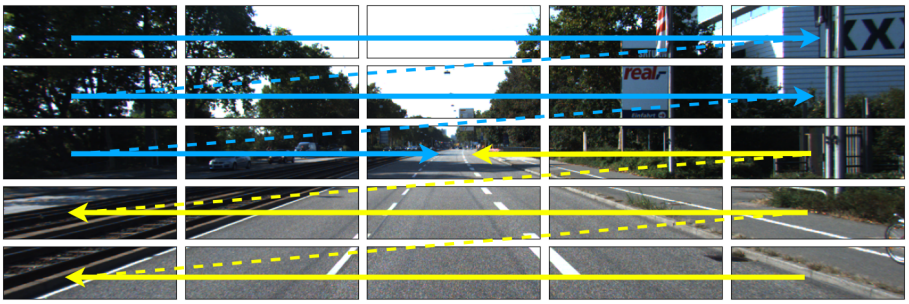
State-space models (SSMs), like other recurrent models, struggle to capture complex contextual information effectively. This limitation stems from the linear time-invariant properties found in Eq. (3). To overcome this challenge, a recent study [16] introduced Mamba, a novel approach that utilizes an input-dependent selection mechanism. This mechanism enhances the model's ability to capture context while maintaining efficiency through a linear-time associative scan algorithm. Building on the success of Mamba, [10] introduced Mamba2. This next-generation model leverages semi-separable matrices, resulting in more efficient training and the ability of having larger recurrent state size.

Bidirectional Mamba. To handle the relative position of each pixel in the image, we proposed BiMamba2 (Fig. 2b), a token mixing method that is based on the building block of Mamba2 [10] with two different scan routes Fig. 3 to capture global visual context. Specifically, the input sequence $\mathbf{u} \in \mathbb{R}^{L\times D}$ is linearly projected to $\mathbf{x} \in \mathbb{R}^{L\times D_e}$, $\mathbf{z} \in \mathbb{R}^{L\times D_e}$ with D_e is the expanded dimension of the model and two sets of parameters $(\mathbf{B}, \mathbf{C}, \Delta)$ for two scanning paths. For each direction, $\mathbf{x}$ and corresponding parameter $(\mathbf{B}, \mathbf{C}, \Delta)$ are concatenated and fed to a 1D-conv layer. Then, they are processed by SSMs and a normalized layer. The output of each scan direction is added together to get the final result.

CrossMamba. While Mamba [10,16] offers a promising alternative to attention mechanisms [51] in various research fields and real-world applications. Their ability in cross-modalities and feature fusion remains limited compared to Transformer models. To address this, we introduce CrossMamba, a novel approach for feature fusion based on the state space models (SSMs). As illustrated in Fig. 2c, let's assume we need to fuse the feature between two sequences $\mathbf{u}_1 \in \mathbb{R}^{L\times D_1}$ and $\mathbf{u}_2 \in \mathbb{R}^{L\times D_2}$. Instead of projecting $\mathbf{u}_1$ onto the parameter set $(\mathbf{B}, \mathbf{C}, \Delta)$, we linearly project $\mathbf{u}_2$ to serve as the input-dependence selection mechanism.



(a) Original Mamba



(b) Bidirectional Mamba

Fig. 3: Illustration of scan methods. (a) The standard Mamba approach scans the image from top-left to bottom-right pixels (**blue routes**), which limits its ability to capture spatial relationships between them. (b) We utilize another scan path (**yellow routes**) to help the model understand the relative position of each pixel in the image. This allows Mamba to capture long-range visual context more effectively.

This projection approach allows the model to select important features of the sequence $\mathbf{u}_1$ based on the information in $\mathbf{u}_2$ and enables the model to fuse them. In this work, to drive visual data, we also follow BiMamba2 and integrate this cross-input dependent selection mechanism with a bidirectional scan route to the proposed CrossBiMamba2. The detailed design can be found in Fig. 2c.

Mamba Encoder-Decoder Architecture. Follow previous design on Encoder Decoder architecture using Transformer [51], we proposed Mamba Encoder Decoder architecture (Fig. 2a). As shown in Fig. 2a, we change the token mixing approach from a self-attention mechanism to BiMamba2 in the encoder and decoder to enhance the efficiency and ability of two scanning routes for visual context. In the decoder, we found that CrossBiMamba2 can be an alternative method to CrossAttention [51] on feature fusion between the query and encode sequence. To integrate context and depth-aware features, we enhance this architecture as a replacement approach for Depth-Aware Transformer [21]. Specifically, the context-aware feature is the input of the encoder with the BiMamba2

method to extract the visual context, while the depth-aware feature is fed into the decoder with another BiMamba2 block. In the end, these two features are fused by the CrossBiMamba2 to enrich the visual context through the depth maps.

3.4 Multi-scale Feature Enhancement

Multi-scale Feature for Depth Prediction. Due to the absence of depth cues, previous works [3, 21, 60] directly predict the depth map from the extracted feature of the vision backbone and use this to guide the learning process. However, these works merely apply several convolutional layers to predict depth maps, which can lead to inaccurate depth hints for the model due to the complexity of depth estimation. Recently, several works [1, 42] show that depth estimation from front view image can benefit from multi-scale features. Being inspired, we leverage Atrous Spatial Pyramid Pooling (ASPP) [6] to extract features at various scales. This allows our model to generate more accurate depth maps by exploiting the benefits of multi-scale information.

Depth-Aware Feature Enhancement. To generate the depth-aware feature, we introduced MSDFE (Fig. 4), a lightweight module leveraging an auxiliary depth estimation task and treating it as a classification problem [14, 40]. As illustrated in Fig. 4, given the input features from the backbone, we utilize ASPP [6] to enhance the benefits of multi-scale features and a 1×1 convolution layers to predict the probability of discretized depth bins. The probability represents the confidence of the depth value concerning each depth bin. To discretize the depth ground truth from LiDAR, we utilize linear-increasing discretization (LID) [40, 49] to formulate the depth bins by Eq. (6):

$$d_c = d_{min} + \frac{d_{max} - d_{min}}{D \times (D+1)} \times d_i \times (d_i + 1) \tag{6}$$

where d_c is the continuous depth value, $[d_{min}, d_{max}]$ is the full depth range to be discretized, D is the number of depth bins and d_i is the depth bin index.

Finally, followed [21], we adopted a group convolution to merge adjacent depth bins and utilize [59] approach to enrich the input feature map and create the final depth aware feature by aggregating the important depth features.

4 Experiments

4.1 Settings

Dataset. We assessed the performance of our MonoDSSMs on the KITTI 3D dataset [15], a widely recognized benchmark for 3D object detection with 7481 images for training and 7581 images for testing. We follow [7] to divide training samples into the training set (3712) and the validation set (3769). We conducted ablation studies using this split to analyze the impact of different components of our MonoDSSMs.

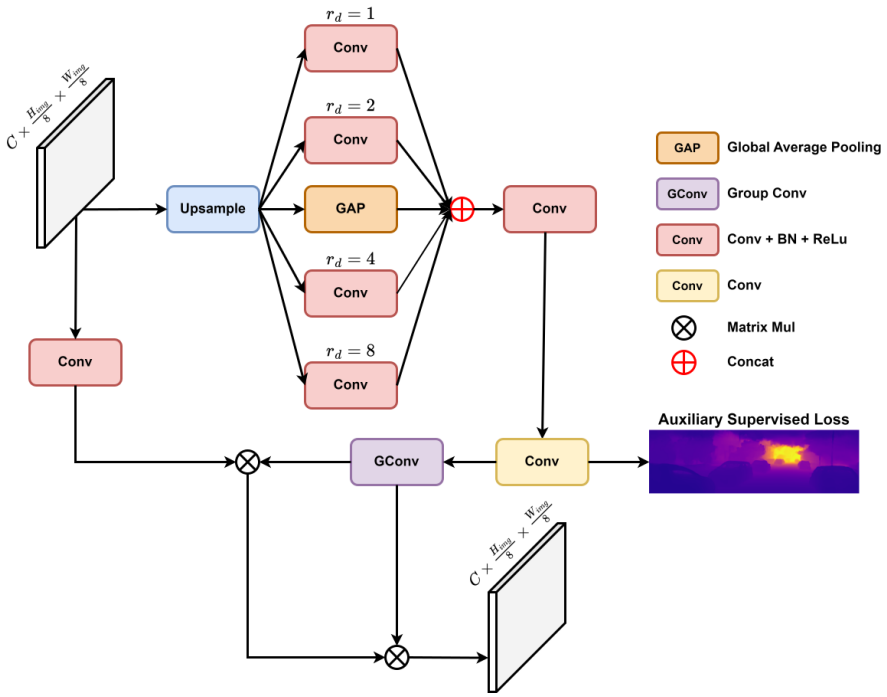


Fig. 4: The proposed Multi-scale depth aware feature enhancement (MS-DFE). The input feature from the backbone is first upsampled to match the resolution of the discretized depth map. Then it is fed to the ASPP module and 1×1 convolution layers to predict the probability of discretized depth bins. A group-convolution layer is then applied to the predicted depth map and fused with the input feature map to produce the final depth-aware feature.

Evaluation metrics. We report detection results for three difficulty levels: easy, moderate, and hard. Evaluation is based on the average precision (AP) of 3D bounding boxes and their corresponding bird's-eye view (BEV) projections. These are denoted as AP_{3D} and AP_{BEV}, respectively, and are calculated at 40 recall positions as suggested by Simonelli *et al.* [46].

Implementation Details. Our network was trained for 120 epochs using the Adam optimizer with a batch size of 12 images on a single NVIDIA 3090 GPU. The learning rate started at 0.0001 and gradually decreased throughout training using a cosine annealing schedule. We leveraged techniques from previous work [21] for anchor box generation, incorporating 3D information from the training data to improve accuracy. Only the top 100 pixels of each image were analyzed to enhance speed during prediction. We normalize all images to a standard resolution of 288×1280. Finally, to refine the detections, we applied a confidence score threshold of 0.75 and Non-Maximum Suppression (NMS) with an IoU threshold of 0.4 to remove redundant bounding boxes.

Table 1: Detection performance of Car category on the KITTI 3D dataset. **Red** numbers indicates the best results for specific metrics, while **blue** denotes the second-best ones. All FPS values were obtained through individual speed tests conducted on each model using a single NVIDIA GeForce RTX 3090 GPU.

Method	FPS	Test, AP_{3D}			Test, AP_{BEV}			Val, AP_{3D}		
		Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
SMOKE [35]	17.6	14.03	9.76	7.84	20.83	14.49	12.75	14.76	12.85	11.50
MonoPair [8]	_	13.04	9.99	8.65	19.28	14.83	12.89	16.28	12.30	10.42
RTM3D [26]	23.3	13.61	10.09	8.18	_	_	_	19.47	16.29	15.57
Kinematic3D [2]	15.6	19.07	12.72	9.17	26.69	17.52	13.10	19.76	14.10	10.47
MonoRUn [5]	20.2	19.65	12.30	10.58	27.94	17.34	15.24	20.02	14.65	12.61
CaDDN [40]	_	19.17	13.41	11.46	27.94	18.91	17.19	23.57	16.31	13.84
PGD [53]	23.9	19.05	11.76	9.39	26.89	16.51	13.49	19.27	13.23	10.65
MonoDLE [37]	25.0	17.23	12.26	10.29	24.79	18.89	16.00	17.45	13.66	11.68
MonoRCNN [45]	24.7	18.36	12.65	10.03	25.48	18.11	14.10	16.61	13.19	10.65
MonoFlex [61]	27.9	19.94	13.89	12.07	28.23	19.75	16.89	23.64	17.51	14.83
GUPNet [36]	28.7	20.11	14.20	11.77	_	_	_	22.76	16.46	13.72
MonoGround [39]	28.2	21.37	14.36	**12.62**	**30.07**	**20.47**	**17.74**	25.24	18.69	15.58
MonoDTR [21]	27.0	**21.99**	**15.39**	**12.73**	28.59	**20.38**	**17.14**	24.52	18.57	15.51
MonoDSSMs-M	**34.5**	19.80	14.15	11.56	28.29	19.59	16.34	**27.10**	**19.10**	**15.84**
MonoDSSMs-A	**34.7**	**21.47**	**14.55**	11.78	**28.84**	19.54	16.30	**26.62**	**18.95**	**15.61**

Design Architecture. We introduce two variants of MonoDSSMs based on the pipeline outlined in Fig. 1. MonoDSSMs-M utilizes our proposed Cross-Mamba module for feature fusion within the Mamba-based encoder-decoder architecture (Fig. 2a). Meanwhile, MonoDSSMs-A leverages CrossAttention [51] as the fusion module in the decoder. We will compare the performance of these two models against previous works.

4.2 Quantitative Results

Experiment on the Car category of KITTI 3D test set. As shown in Tab. 1, we compare our MonoDSSMs-M and MonoDSSMs-A with several state-of-the-art monocular 3D object detection methods on the KITTI test set. It can be observed that our approach achieves competitive performance compared with other methods. However, our proposed model does not achieve outstanding results compared with the baseline, this can be explained by the trade-off between accuracy and efficiency. Our setting MonoDSSMs-A only reduces about 2.4%, 5.5%, and 7.5% on the easy, moderate, and hard levels respectively for the AP_{3D} metric. However, for the AP_{BEV} metric, our model is even better at the easy level and only reduces 4.1% and 5% on the remaining 2 levels. Meanwhile, the inference speed is increased by **28.5%** compared to the baseline.

Experiment on the Car category of KITTI 3D val set. We also evaluated our approach on the KITTI validation dataset using the AP_{3D} metrics as listed in Tab. 1. Our method outperforms previous works thanks to

the Mamba-based encoder-decoder architecture and the multi-scale depth enhancement strategy. Specifically, compared to the baselines, MonoDTR [21], MonoDSSMs-M achieves significant improvements in AP_{3D} at the 0.7 IoU threshold across three settings: **2.58**/**0.53**/**0.33**. Similarly, MonoDSSMs-A shows improvements **2.10**/**0.38**/**0.10**.

Efficiency analysis. We evaluated the speed of our models by processing the entire KITTI validation dataset on a single NVIDIA 3090 GPU. The results, shown in Tab. 1 demonstrate that our models achieve real-time performance at *34* FPS, signifying the efficiency of our approach. Furthermore, our MonoDSSMs are significantly faster than previous state-of-the-art monocular 3D object detection methods, running at speeds **1.23x** and **1.28x** than MonoGround [39] and MonoDTR [21], respectively.

4.3 Ablation Study

Model Efficiency. In Tab. 2, we first change the backbone of the baseline from DLA-102 [58] to DLA-34 (b). As mentioned in the previous section, the performance of our approach drops 2–7% in 3 different levels of the evaluation metric. However, the model has 2.3x fewer parameters and 1.33x faster computation time, exceeds the threshold of 30 fps, which is both efficient and applicable to real-world applications.

Effectiveness of Mamba. In Tab. 2, we conduct various experiments with different settings: (c) Replacing SelfAttention in the encoder and decoder of Depth-Aware Transformer [21] with Mamba, (d) Replacing the original Mamba with Bidirection Mamba to enhance the visual context. Firstly, we see that using Mamba (c) provides a modest performance improvement without sacrificing efficiency. Furthermore, Bidirection-Mamba (d) offers a significant boost in detection accuracy with minimal computational overhead. These results highlight the effectiveness of Mamba, particularly the novel Bidirection-Mamba architecture, in handling visual data.

Multi scale depth prediction. To investigate the impact of multi-scale depth map prediction on detection performance, we conducted an additional experiment (e) as shown in Tab. 2. Our results demonstrate that incorporating ASPP outperforms the simpler approach in prior work on the KITTI validation set. Especially, employing various dilation rates during depth map feature extraction significantly boosted performance, particularly for moderate and hard difficulty levels. This improvement likely stems from capturing more accurate depth information for distant objects.

Mamba-based feature fusion. As shown in Tab. 3, we employ MonoDSSMs-A and MonoDSSMs-M to evaluate the effectiveness of the proposed Mamba-based feature fusion. In the KITTI validation set, our experiments demonstrate that the proposed CrossMamba achieves better detection performance than CrossAttention, particularly at easier difficulty levels with a 2% gap. However, the KITTI test set results for MonoDSSMs-M are lower. This might be due to some mismatch between the dataset distribution of the two sets. Despite this, based on the positive results in the KITTI validation set, we believe that the

Table 2: Analysis of different components of our approach on the Car category of the KITTI validation set. 'DLA34' denotes using DLA-34 [58] as backbone. 'S6', 'BiS6' denotes using Mamba2 [10] and the proposed BiMamba2 as an alternative approach to SelfAttention [51]. 'ASPP' denotes the ASPP approach [6] for depth map prediction by utilizing various dilation rates.

	Ablation				FPS	Params	AP_{3D}@IoU=0.7			AP_{BEV}@IoU=0.7		
	DLA-34	S6	BiS6	ASPP			Easy	Mod.	Hard	Easy	Mod.	Hard
(a)					27.0	54.25	24.52	18.57	15.51	33.33	25.35	21.68
(b)	✓				36.0	23.28	24.04	18.02	14.80	33.66	24.09	20.24
(c)	✓	✓			36.2	23.39	25.16	18.33	15.14	33.85	24.65	20.24
(d)	✓		✓		35.5	23.43	26.57	18.59	15.28	34.82	25.04	20.34
(e)	✓		✓	✓	34.7	23.61	26.62	18.95	15.61	35.96	25.90	22.02

Table 3: Comparison of different feature fusion methods on the Car category of the KITTI dataset. 'CA' denotes Cross Attention [51] module as a feature fusion method within the encoder-decoder architecture (MonoDSSMs-A). 'CS6' denotes the proposed CrossMamba as an alternative approach to CA (MonoDSSMs-M).

Ablation		Test, AP_{3D}			Test, AP_{BEV}			Val, AP_{3D}			Val, AP_{BEV}		
CA	CS6	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
✓		21.47	14.55	11.78	28.84	19.54	16.30	26.62	18.95	15.61	35.96	25.90	22.02
	✓	19.80	14.15	11.56	28.29	19.59	16.34	27.10	19.10	15.84	37.73	25.86	22.12

proposed CrossMamba has promise as a powerful feature fusion approach based on State Space Models (SSMs) and can be further developed in future work.

4.4 Qualitative Results

Fig. 5 presents qualitative examples from the KITTI validation set. Compared to the baseline MonoDTR model, our MonoDSSMs model yields predictions that are significantly closer to the ground truth. As we can observe in Fig. 5, our prediction (blue boxes) is very close to the groundtruth (red boxes), while there are still recognizable gaps between the baseline prediction (green boxes) and the groundtruth. For sample (b), we can see that our proposed model can detect the car that the baseline cannot. Still, the car in the bottom-left is hard to recognize due to the occlusion. In sample (c), our model misses one car at the bottom of the BEV. However, it is reasonable since the groundtruth is marked using LiDAR scans. This explains why our model can detect one car (the blue box at the bottom-left corner of the BEV) but it does not appear in the groundtruth due to occlusion in the LiDAR scan. The baseline, in this case, missed that car as well. Overall, the visualization results in Fig. 5 show that our proposed model has immensely improved the detection accuracy from the baseline model.

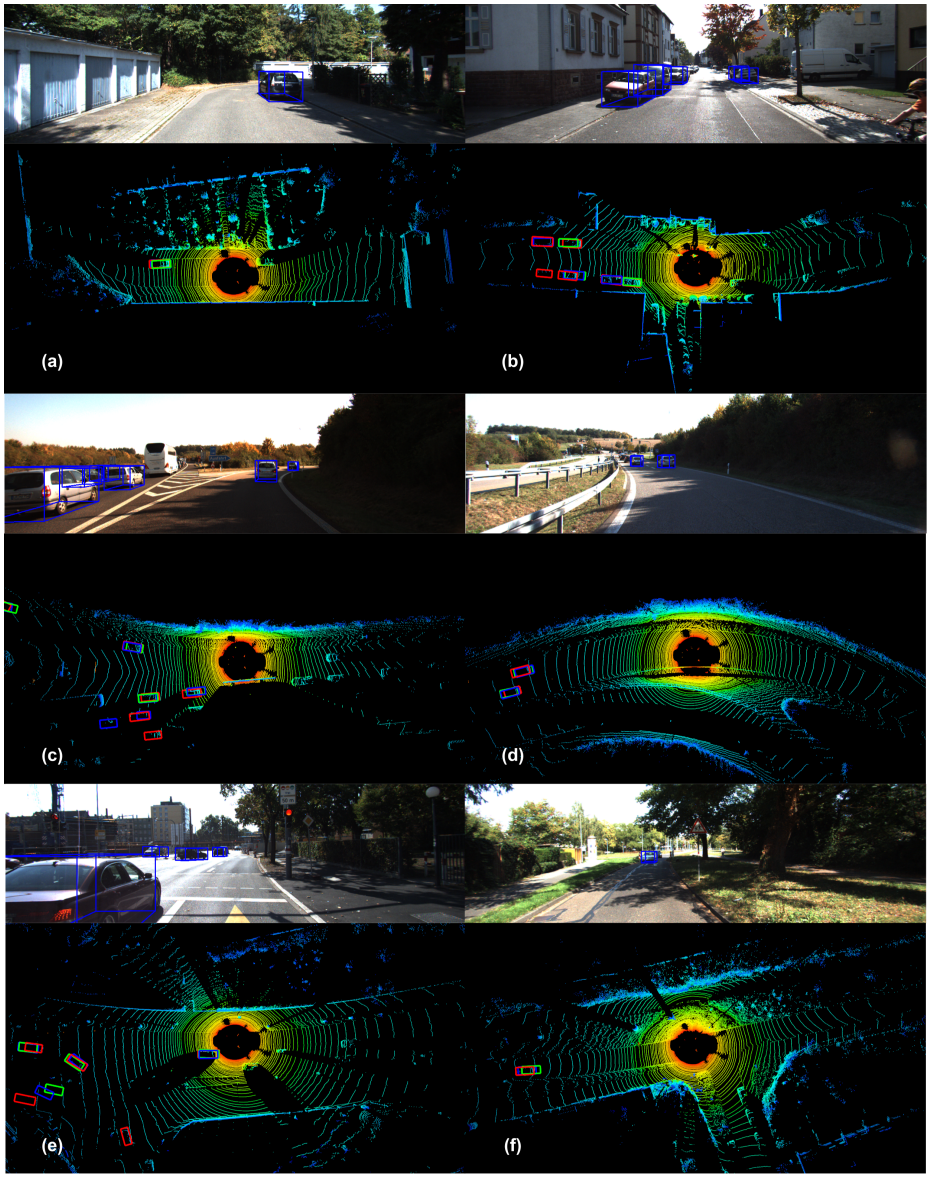


Fig. 5: Qualitative examples on the KITTI validation set. We provide the predictions on both the image view (top) and bird's-eye view (bottom). **Blue** boxes indicate predictions from MonoDSSMs in the image and BEV plane. **Red** boxes represent ground truth, while **green** boxes show predictions from the baseline model on the BEV. For optimal viewing, zoom in and use color.

5 Conclusion

In our work, we present MonoDSSMs, an efficient monocular 3D object detection system with a cutting-edge Mamba-based architecture. The innovative Bi-Mamba2 with bidirectional scan allows the model to capture long-range visual context while maintaining efficiency. Additionally, we introduce CrossMamba, an alternative approach for CrossAttention to integrate depth/context-aware features globally. Our utilization of a multi-scale strategy ensures the production of a higher-quality depth map, enhancing depth cues for our models. Our experiments demonstrate that MonoDSSMs achieve competitive results with significantly faster inference times compared to previous work. MonoDSSMs stand as a formidable Mamba-based baseline for future monocular 3D object detection research.

Limitations. Despite the capability to achieve real-time detections, there is still room for improving the performance of MonoDSSMs. As discussed in Sec. 4.2, the accuracy on the KITTI 3D test set shows a slight decrease due to the efficiency trade-off. Although this small performance drop is acceptable for real-time applications, future work will focus on improving accuracy while maintaining efficiency.

Acknowledgement. We acknowledge Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for supporting this study.

References

1. Agarwal, A., Arora, C.: Depthformer: Multiscale vision transformer for monocular depth estimation with global local information fusion. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 3873–3877 (2022)
2. Brazil, G., Pons-Moll, G., Liu, X., Schiele, B.: Kinematic 3d object detection in monocular video. In: Computer Vision – ECCV 2020. pp. 135–152. Springer International Publishing, Cham (2020)
3. Bui, M.Q.V., Ngo, D.T., Pham, H.A., Nguyen, D.D.: Gac3d: improving monocular 3d object detection with ground-guide model and adaptive convolution. PeerJ Computer Science **7** (2021)
4. Cai, Y., Li, B., Jiao, Z., Li, H., Zeng, X., Wang, X.: Monocular 3d object detection with decoupled structured polygon estimation and height-guided depth estimation. In: AAAI Conference on Artificial Intelligence (2020)
5. Chen, H., Huang, Y., Tian, W., Gao, Z., Xiong, L.: Monorun: Monocular 3d object detection by reconstruction and uncertainty propagation. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10374–10383 (2021)
6. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. ArXiv (2017)
7. Chen, X., Kundu, K., Zhu, Y., Berneshawi, A.G., Ma, H., Fidler, S., Urtasun, R.: 3d object proposals for accurate object class detection. In: Advances in Neural Information Processing Systems. vol. 28. Curran Associates, Inc. (2015)

8. Chen, Y., Tai, L., Sun, K., Li, M.: Monopair: Monocular 3d object detection using pairwise spatial relationships. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12090–12099 (2020)
9. Dao, T., Fu, D.Y., Saab, K.K., Thomas, A.W., Rudra, A., Ré, C.: Hungry hungry hippos: Towards language modeling with state space models. ArXiv (2022)
10. Dao, T., Gu, A.: Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In: Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 10041–10071. PMLR (21–27 Jul 2024)
11. Ding, M., Huo, Y., Yi, H., Wang, Z., Shi, J., Lu, Z., Luo, P.: Learning depth-guided convolutions for monocular 3d object detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11669–11678 (2020)
12. Dong, Z., Ji, H., Huang, X., Zhang, W., Zhan, X., Chen, J.: Pep: a point enhanced painting method for unified point cloud tasks. CoRR (2023)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR (2021)
14. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 2002–2011 (2018)
15. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3354–3361 (2012)
16. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. ArXiv (2023)
17. Gu, A., Goel, K., R'e, C.: Efficiently modeling long sequences with structured state spaces. ArXiv (2021)
18. He, C., Zeng, H., Huang, J., Hua, X.S., Zhang, L.: Structure aware single-stage 3d object detection from point cloud. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11870–11879 (2020)
19. Hu, V.T., Baumann, S.A., Gui, M., Grebenkova, O., Ma, P., Fischer, J., Ommer, B.: Zigma: A dit-style zigzag mamba diffusion model. In: ECCV (2024)
20. Huang, J., Huang, G., Zhu, Z., Du, D.: Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. ArXiv (2021)
21. Huang, K., Wu, T., Su, H., Hsu, W.H.: Monodtr: Monocular 3d object detection with depth-aware transformer. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4002–4011 (2022)
22. Huang, T., Pei, X., You, S., Wang, F., Qian, C., Xu, C.: Localmamba: Visual state space model with windowed selective scan. ArXiv (2024)
23. Kumar, A., Brazil, G., Corona, E., Parchami, A., Liu, X.: DEVIANT: depth equivariant network for monocular 3d object detection. In: Computer Vision - ECCV 2022 - 17th European Conference, Proceedings, Part IX. Lecture Notes in Computer Science, vol. 13669, pp. 664–683. Springer (2022)
24. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12689–12697 (2019)
25. Li, K., Li, X., Wang, Y., He, Y., Wang, Y., Wang, L., Qiao, Y.: Videomamba: State space model for efficient video understanding. In: ECCV (2024)

26. Li, P., Zhao, H., Liu, P., Cao, F.: RTM3D: real-time monocular 3d detection from object keypoints for autonomous driving. In: Computer Vision - ECCV 2020 - 16th European Conference, Proceedings, Part III. Lecture Notes in Computer Science, vol. 12348, pp. 644–660. Springer (2020)
27. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Yu, Q., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. ArXiv (2022)
28. Li, Z., Qu, Z., Zhou, Y., Liu, J., Wang, H., Jiang, L.: Diversity matters: Fully exploiting depth clues for reliable monocular 3d object detection. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2781–2790 (2022)
29. Liang, D., Zhou, X., Wang, X., Zhu, X., Xu, W., Zou, Z., Ye, X., Bai, X.: Pointmamba: A simple state space model for point cloud analysis. ArXiv (2024)
30. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: Computer Vision – ECCV 2016. pp. 21–37. Springer International Publishing, Cham (2016)
31. Liu, X., Xue, N., Wu, T.: Learning auxiliary monocular contexts helps monocular 3d object detection. In: AAAI Conference on Artificial Intelligence (2021)
32. Liu, Y., Wang, T., Zhang, X., Sun, J.: Petr: Position embedding transformation for multi-view 3d object detection. ArXiv (2022)
33. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Liu, Y.: Vmamba: Visual state space model. ArXiV (2024)
34. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 9992–10002 (2021)
35. Liu, Z., Wu, Z., T'oth, R.: Smoke: Single-stage monocular 3d object detection via keypoint estimation. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp. 4289–4298 (2020)
36. Lu, Y., Ma, X., Yang, L., Zhang, T., Liu, Y., Chu, Q., Yan, J., Ouyang, W.: Geometry uncertainty projection network for monocular 3d object detection. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3091–3101 (2021)
37. Ma, X., Zhang, Y., Xu, D., Zhou, D., Yi, S., Li, H., Ouyang, W.: Delving into localization errors for monocular 3d object detection. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4719–4728 (2021)
38. Mehta, H., Gupta, A., Cutkosky, A., Neyshabur, B.: Long range language modeling via gated state spaces. CoRR (2022)
39. Qin, Z., Li, X.: Monoground: Detecting monocular 3d objects from the ground. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3783–3792 (2022)
40. Reading, C., Harakeh, A., Chae, J., Waslander, S.L.: Categorical depth distribution network for monocular 3d object detection. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 8551–8560 (2021)
41. Redmon, J., Divvala, S.K., Girshick, R.B., Farhadi, A.: You only look once: Unified, real-time object detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 779–788 (2015)
42. Sagar, A.: Monocular depth estimation using multi scale neural network and feature fusion. 2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW) pp. 656–662 (2020)

43. Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., Li, H.: Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10526–10535 (2020)
44. Shi, S., Wang, X., Li, H.: Pointrcnn: 3d object proposal generation and detection from point cloud. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 770–779 (2018)
45. Shi, X., Ye, Q., Chen, X., Chen, C., Chen, Z., Kim, T.K.: Geometry-based distance decomposition for monocular 3d object detection. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 15152–15161 (2021)
46. Simonelli, A., Bulò, S.R., Porzi, L., López-Antequera, M., Kontschieder, P.: Disentangling monocular 3d object detection. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 1991–1999 (2019)
47. Smith, J., Warrington, A., Linderman, S.W.: Simplified state space layers for sequence modeling. ArXiv (2022)
48. Su, H., Jampani, V., Sun, D., Gallo, O., Learned-Miller, E.G., Kautz, J.: Pixel-adaptive convolutional neural networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 11158–11167 (2019)
49. Tang, Y., Dorn, S., Savani, C.: Center3d: Center-based monocular 3d object detection with joint depth understanding. In: Pattern Recognition. pp. 289–302. Springer International Publishing, Cham (2021)
50. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: Proceedings of the 38th International Conference on Machine Learning, ICML. Proceedings of Machine Learning Research, vol. 139, pp. 10347–10357. PMLR (2021)
51. Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Neural Information Processing Systems (2017)
52. Wang, L., Du, L., Ye, X., Fu, Y., Guo, G., Xue, X., Feng, J., Zhang, L.: Depth-conditioned dynamic message propagation for monocular 3d object detection. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 454–463 (2021)
53. Wang, T., Zhu, X., Pang, J., Lin, D.: Probabilistic and geometric depth: Detecting objects in perspective. In: Conference on Robot Learning (2021)
54. Wang, Y., Guizilini, V.C., Zhang, T., Wang, Y., Zhao, H., Solomon, J.: Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. ArXiv (2021)
55. Xiong, K., Zhang, D., Liang, D., Liu, Z., Yang, H., Dikubab, W., Cheng, J., Bai, X.: You only look bottom-up for monocular 3d object detection. IEEE Robotics and Automation Letters **8**(11), 7464–7471 (2023)
56. Yang, C., Chen, Y., Tian, H., Tao, C., Zhu, X., Zhang, Z., Huang, G., Li, H., Qiao, Y., Lu, L., Zhou, J., Dai, J.: Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 17830–17839 (2022)
57. Yin, T., Zhou, X., Krähenbühl, P.: Center-based 3d object detection and tracking. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 11779–11788 (2020)
58. Yu, F., Wang, D., Darrell, T.: Deep layer aggregation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 2403–2412 (2017)
59. Zhang, F., Chen, Y., Li, Z., Hong, Z., Liu, J., Ma, F., Han, J., Ding, E.: Acfnet: Attentional class feature network for semantic segmentation. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 6797–6806 (2019)

60. Zhang, R., Qiu, H., Wang, T., Guo, Z., Cui, Z., Xu, X., Qiao, Y.J., Gao, P., Li, H.: Monodetr: Depth-guided transformer for monocular 3d object detection. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 9121–9132 (2022)
61. Zhang, Y., Lu, J., Zhou, J.: Objects are different: Flexible monocular 3d object detection. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3288–3297 (2021)
62. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. ArXiv (2019)
63. Zhou, Y., Tuzel, O.: Voxelnet: End-to-end learning for point cloud based 3d object detection. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4490–4499. IEEE Computer Society (2018)
64. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. ArXiv (2024)
65. Zhu, M., Ge, L., Wang, P., Peng, H.: Monoedge: Monocular 3d object detection using local perspectives. 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) pp. 643–652 (2023)