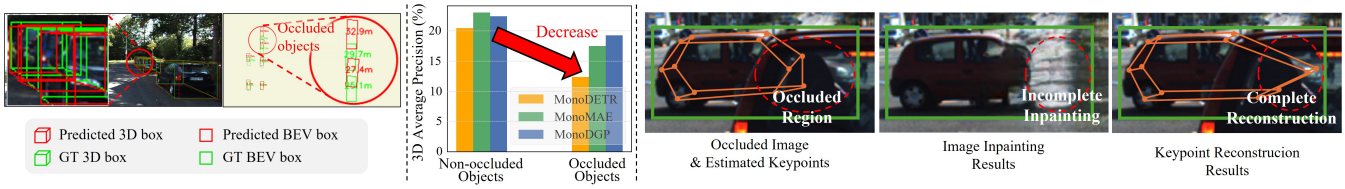
2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

MonoKey: Occlusion-Robust Keypoint-based Monocular 3D Object Detection with Prior Guidance

Yeon Woo Cho†, Jung Woo Cheon†, Jae Hyun Yoon and Seok Bong Yoo*



(a) (b) (c)

Fig. 1: (a) Visual detection results of a state-of-the-art monocular 3D object detection model (MonoDGP) under occlusion. (b) Comparison of detection performance among existing methods for occluded and non-occluded objects. (c) Comparison of image inpainting using LaMa [1] and keypoint reconstruction approaches.

***Abstract*— Monocular 3D object detection has garnered attention due to its cost-efficiency and simpler setup compared with multisensor systems. In this task, an accurate depth estimation is crucial for precise object localization, however extracting sufficient depth cues from a single image remains challenging. Moreover, when occlusions occur, structural cues become limited, making precise object localization increasingly difficult. To address these problems, we propose MonoKey, a keypoint-based monocular 3D object detection method that is robust to occlusion. MonoKey applies 2D keypoints due to their suitability for recovering occluded regions. The occlusion-robust 2D keypoint detection approach estimates object keypoints and reconstructs occluded ones using prior information. The frequency-based global-local depth predictor estimates 3D cues using fast Fourier convolution to incorporate global and local contexts. These 3D cues and keypoints are fused in a 3D detection decoder. Relational graph refinement adjusts the initial bounding boxes for improved localization. The experimental results indicate that MonoKey outperforms the existing monocular 3D object detection methods. The source code is available at https://github.com/yeonwoo29/MonoKey.git.**

I. INTRODUCTION

Three-dimensional (3D) object detection is an essential task in robotic automation, such as autonomous driving. Approaches that apply multiple sensors, such as LiDAR or stereo cameras, can measure object depth and enable precise localization. However, these methods require accurate calibration between sensors and involve excessive costs for sensor configuration, limiting their scalability and widespread adoption. In contrast, monocular 3D object detection estimates object locations using only a single red, green, and blue (RGB) camera. This approach eliminates the need for multisensor calibration and reduces the system cost. These advantages have attracted significant research interest in monocular-based approaches. This strategy can also benefit edge deployment scenarios, including those in autonomous vehicles, mobile robotics, augmented/virtual reality systems, and uncrewed aerial vehicles, where low-cost, lightweight perception modules are crucial.

Despite these advantages, monocular 3D object detection still suffers from the difficulty of accurately estimating depth from a single view. This limitation becomes more pronounced under occlusion, where partial object visibility causes severe information loss. Monocular methods rely on RGB images, which do not utilize the geometric layout of an object; thus, many state-of-the-art (SOTA) approaches, including MonoDETR [2] and MonoDGP [3], depend heavily on appearance-based cues, including color, texture, or object boundaries. These cues are often corrupted or missing when objects are occluded, degrading performance. Figure 1(a) illustrates this problem, where the predicted 3D bounding boxes under occlusion deviate significantly from the ground truth (GT). Figure 1(b) illustrates the decline in detection performance between occluded and non-occluded cases.

To overcome these challenges, we propose MonoKey, a keypoint-guided monocular 3D object detection framework designed to be robust under occlusion. Instead of relying solely on appearance information, MonoKey structurally infers occluded regions by introducing 12 semantic keypoints that capture the geometric layout of an object. Represented as (x, y) coordinates, these keypoints form a compact 24-dimensional vector, reducing feature dimensionality compared to dense image representations, lowering computational complexity. Occluded keypoints are reconstructed via

This work was supported by Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (MOTIE) (P0020536) and the IITP grant funded by the Korea Government (MSIT) (RS-2024–00437718, RS-2023–00256629, RS-2022–00156287).

The authors are with Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju 61186, South Korea.
{cyw2628@jnu.ac.kr, cjwdada@jnu.ac.kr, jhyoon@gmail.com, sbyoo@jnu.ac.kr}
† These authors contributed equally to this work.
* Corresponding author: Seok Bong Yoo

an autoencoder that incorporates coarse yaw and structural symmetry as priors, while applying symmetry constraints and yaw consistency to reduce uncertainty during reconstruction.

Figure 1(c) compares the proposed with the image pixel-based inpainting baseline (LaMa [1]). Although inpainting fails to reconstruct the object structure in the GT bounding box, the proposed keypoint-based method succeeds by applying structural cues. By combining semantic keypoints, geometric priors, and global depth features, MonoKey is robust against occlusion while maintaining low-cost monocular input requirements.

Furthermore, MonoKey integrates a frequency-based global-local depth predictor (FDP) to cope with the loss of structural details under occlusion. This module balances low- and high-frequency components by applying frequency-domain representations, enabling a more accurate depth estimation in partially occluded scenes. A relational graph refinement module mitigates the effects of occlusion by modeling the relationships between the object keypoints and initial bounding boxes, enhancing localization precision.

The primary contributions are summarized as follows:

- We propose an occlusion-robust 2D keypoint detection method that localizes visible keypoints and reconstructs occluded ones via a prior-guided autoencoder, mitigating the loss of structural information under occlusion.
- We propose an FDP that mitigates the influence of occlusion on depth estimation by capturing global context and fine spatial details in the frequency domain. Balancing low- and high-frequency components alleviates spectral bias and improves depth prediction robustness in partially occluded scenes.
- We propose a relational graph-based module that applies keypoints and yaw information to model spatial relationships between objects and refine 3D bounding boxes via graph reasoning, thereby improving the localization of occluded objects by using the accurate detection of non-occluded ones.

II. RELATED WORK

A. Monocular 3D Object Detection

Monocular 3D object detection using RGB images alone has garnered significant attention due to its low cost. These methods [2], [4]–[18] infer 3D attributes, such as position, size, and yaw, directly from a single image. Although efficient, these methods lack the complementary cues that allow the detector to focus precisely on the object. In response to the challenges of monocular perception, some methods have incorporated auxiliary signals, such as generating bird's-eye view representations or using 2D corner projections [19], [20]. The proposed approach follows this paradigm but employs 2D keypoint information as an auxiliary signal to overcome the limitations of purely image-based approaches.

Occlusion is a significant source of error in 3D object detection. Although early methods have often assumed full object visibility, recent approaches have addressed occlusion from diverse perspectives. [21] An occlusion partial point cloud augmentation technique [22] improves robustness by synthetically occluding objects. However, this method relies on data augmentation. RGB-based methods such as M3D-RPN [23] incorporate occlusion reasoning in its proposal stage, but still struggles with complex occlusion patterns in real-world scenarios. Moreover, MonoMAE [24] incorporates occlusion-aware masked autoencoding to handle occluded regions. In contrast, our method explicitly reconstructs missing geometric structures using semantic keypoints and prior-guided autoencoding. By leveraging object keypoints to capture the underlying geometric layout, our approach enables accurate 3D box reconstruction even under severe occlusion. This work employs semantic keypoints that represent rigid object geometry (e.g., vehicles) for monocular 3D detection, beyond their prior use in datasets such as CarFusion [25] or SKoPe [26].

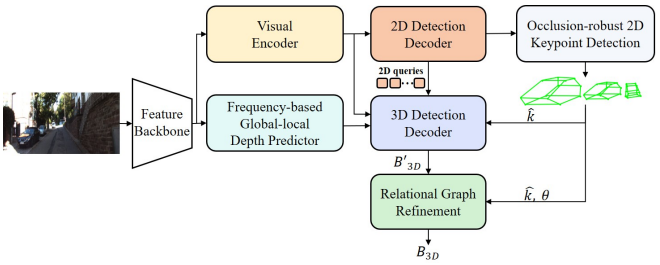


Fig. 2: Overall architecture of MonoKey framework.

B. 2D Object Keypoint Detection

Keypoint-based methods are robust to background variations and focus on object structure. These methods have been widely studied under the unified task of keypoint-based detection [27]–[37]. In driving scenarios, object keypoint datasets [25], [26] structurally define vehicle keypoints. Generally, for non-rigid objects such as humans, keypoints represent the pose. In contrast, for rigid objects such as vehicles, keypoints represent the geometric layout. Therefore, this work utilizes keypoints that represent the geometric layout to perform 3D rigid object detection.

III. METHOD

A. Overview

Figure 2 illustrates the proposed MonoKey, a monocular 3D object detection framework designed to handle occlusion. MonoKey extracts visual features with a backbone and employs an occlusion-robust 2D keypoint detection module to estimate visible keypoints $\hat{k}$. The estimated keypoints are embedded using spatial positional encodings and provided as input to the 3D detection decoder. In parallel, a FDP applies frequency decomposition to generate depth features capturing both local details and global context, which are also provided as input to the 3D detection decoder. The decoder applies cross-attention to these inputs to generate the initial 3D bounding boxes B'_{3D}, which are refined by a relational graph module modeling spatial dependencies to resolve overlaps and improve yaw accuracy, producing the final 3D bounding boxes B_{3D}.

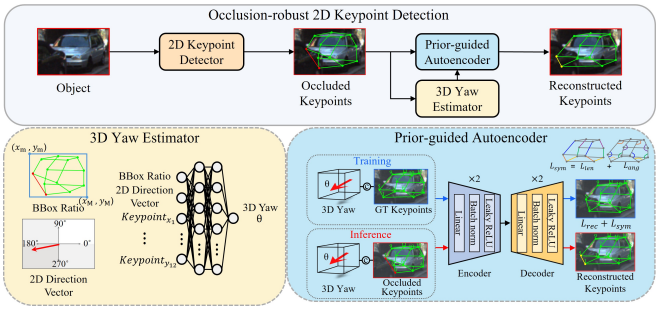


Fig. 3: Illustration of the occlusion-robust 2D keypoint detection.

B. Occlusion-robust 2D Keypoint Detection

1) Domain Adaptive 2D Keypoint Detection: This work introduces object-level keypoints and addresses the domain gap (e.g., viewpoint, background, and resolution) between the CarFusion dataset [25] and real driving environments to enable effective 2D keypoint detection. The CarFusion dataset is used for training, and 2D keypoint detection is performed on object-level crops with domain adaptation.

In the initial block of Fig. 2, object bounding boxes are localized using a 2D detector, and keypoints are detected in each cropped region corresponding to the detected boxes. For vehicle objects, we establish a dedicated 2D keypoint detection framework by defining a set of meaningful object keypoints and training the keypoint detector. The object and keypoint detectors are implemented using CenterNet [38] with a DLA-34 [39] backbone. CenterNet predicts object centers and regresses keypoints from shared backbone features. This work applies a method that minimizes the maximum mean discrepancy (MMD) [40] between the source and target domain feature distributions to reduce the domain gap. The MMD loss is defined as follows:

$$\mathcal{L}_{\text{mmd}}^{\text{det}} = \left\| \frac{1}{n^s} \sum_{p=1}^{n^s} f_p^s - \frac{1}{n^t} \sum_{q=1}^{n^t} f_q^t \right\|_2, \tag{1}$$

where f_p^s and f_q^t denote the feature embeddings from the source and target domains, respectively, and n^s and n^t represent the batch sizes for the source and target samples.

2) 3D Yaw Estimator: The 3D yaw angle θ, which serves as a global geometric prior for keypoint reconstruction, is estimated to complement the partial keypoint observations under occlusion. This global cue provides coarse directional information regarding the vehicle structure and guides the autoencoder to produce structurally consistent keypoints.

As illustrated in Fig. 3, the 3D yaw θ is estimated using a multilayer perceptron (MLP)–based yaw estimator, which takes the 2D keypoints detected in the previous step as input. The input comprises 12 semantic keypoints (each with 2D coordinates, forming a 24-dimensional vector), a 2D direction vector (four dimensions), and the height-to-width ratio of the 2D bounding box (one dimension). The 12 keypoints are predefined to represent the front and rear wheels, indicating the headlights, and to mark the rooftop corners. The direction vector is computed as the mean *(x, y)* coordinates of six front-facing keypoints and the six rear-facing keypoints. The bounding box ratio serves as a geometric prior that correlates with yaw. Resulting in a 29-dimensional input to the MLP, comprising two hidden layers with 128 units and a sigmoid output layer that predicts θ.

3) Prior-guided Autoencoder: The predicted 3D yaw angle θ serves as a global geometric prior for reconstructing occluded keypoints. We propose an autoencoder that combines global and local geometric priors. The network consists of an encoder and a decoder, each with two fully connected layers, batch normalization, and a LeakyReLU activation.

In the training stage (blue arrows in Fig. 3), the CarFusion dataset provides the GT 2D coordinates of 12 keypoints concatenated with the yaw θ as input, and the network is trained to minimize the difference between the input and the reconstructed output. In the inference stage (red arrows), the 12 partially observed 2D keypoints under occlusion concatenated with the predicted yaw θ are used as the model input. Since the model is trained only with GT keypoints, it reconstructs occluded ones based on learned structural patterns. θ provides coarse pose cues, and symmetry-based local geometric priors based on edge lengths and angles consistently observed across the dataset further enhance the reconstruction performance. The loss function is defined as:

$$\mathcal{L}^{\text{ae}} = \mathcal{L}_{\text{rec}}^{\text{ae}} + \lambda_{\text{sym}} \mathcal{L}_{\text{sym}}^{\text{ae}}, \tag{2}$$

where $\mathcal{L}_{\text{rec}}^{\text{ae}}$ is the reconstruction loss, $\mathcal{L}_{\text{sym}}^{\text{ae}}$ is the structural symmetry loss, and λ_{sym} balances the two terms.

The reconstruction loss minimizes the discrepancy between the predicted and input 2D coordinates, encouraging the autoencoder to recover a coherent set of keypoints from partial observations. It is defined as the mean Euclidean distance between the predicted and input 2D coordinates:

$$\mathcal{L}_{\text{rec}}^{\text{ae}} = \frac{1}{12} \sum_{i=1}^{12} \left\| \hat{k}_i - k_i \right\|_2, \tag{3}$$

where $\hat{k}_i$ and k_i denote the reconstructed and input 2D coordinates of the i-th keypoint, respectively.

The symmetry loss enforces geometric consistency with edge-length and angle terms:

$$\mathcal{L}_{\text{sym}}^{\text{ae}} = \mathcal{L}_{\text{len}}^{\text{ae}} + \mathcal{L}_{\text{ang}}^{\text{ae}}, \tag{4}$$

where $\mathcal{L}_{\text{len}}^{\text{ae}}$ penalizes asymmetric edge lengths and $\mathcal{L}_{\text{ang}}^{\text{ae}}$ penalizes inconsistent angles between symmetric keypoint triplets. Both terms are normalized to [0, 1], requiring no additional weighting.

The edge-length symmetry loss $\mathcal{P}_{\text{len}}$ is computed over a predefined set of symmetric edge pairs $\mathcal{P}_{\text{len}}$:

$$\mathcal{L}_{\text{len}}^{\text{ae}} = \frac{1}{|\mathcal{P}_{\text{len}}|} \sum_{(a,b;\,c,d) \in \mathcal{P}_{\text{len}}} \left(\frac{\Delta_{ab,cd}}{Z_{ab,cd}} \right)^2, \tag{5}$$

$$\Delta_{ab,cd} = \left\| \hat{k}_a - \hat{k}_b \right\|_2 - \left\| \hat{k}_c - \hat{k}_d \right\|_2, \tag{6}$$

$$Z_{ab,cd} = \max \left(\left\| \hat{k}_a - \hat{k}_b \right\|_2, \left\| \hat{k}_c - \hat{k}_d \right\|_2 \right) + \varepsilon, \tag{7}$$

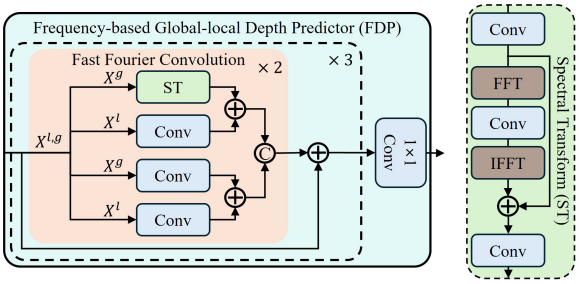


Fig. 4: Illustration of the frequency-based global-local depth predictor.

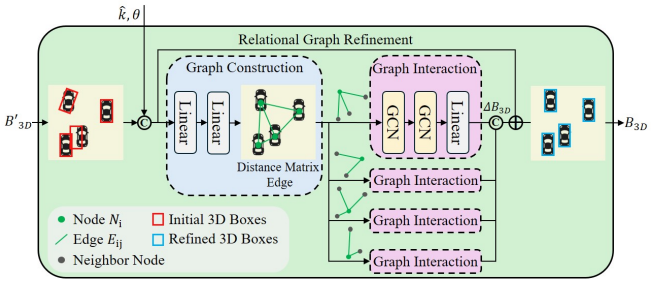


Fig. 5: Illustration of the relational graph refinement.

where ε is a small constant to avoid division by zero.

The angle symmetry loss $\mathcal{P}_{\text{ang}}$ is computed over a predefined set of symmetric triplet pairs $\mathcal{P}_{\text{ang}}$:

$$\mathcal{L}_{\text{ang}}^{\text{ae}} = \frac{1}{|\mathcal{P}_{\text{ang}}|} \sum_{(a,b,c;\, d,e,f)\in\mathcal{P}_{\text{ang}}} \left(\frac{\Delta_{abc,def}^{\angle}}{2} \right)^2, \tag{8}$$

$$\Delta_{abc,def}^{\angle} = \cos\angle(\hat{k}_a, \hat{k}_b, \hat{k}_c) - \cos\angle(\hat{k}_d, \hat{k}_e, \hat{k}_f), \tag{9}$$

$$\cos\angle(u, v, w) = \frac{(u-v)\cdot(w-v)}{\|u-v\|_2 \|w-v\|_2}. \tag{10}$$

These priors, based on edge lengths and angles, enable robust reconstruction from occluded keypoints.

C. Frequency-based Global-local Depth Predictor

Existing monocular 3D object detectors derived from [2], which estimate depth without extra sensor data, have garnered attention due to their cost-efficiency. However, depth estimation methods based on the convolutional neural network suffer from spectral bias, where neural networks favor low-frequency components over high-frequency ones. This bias causes blurry predictions and loss of fine details, especially under occlusion where high-frequency cues (e.g., edges, or boundaries) are crucial. This work adopts an FDP to address this, which processes feature representations in the frequency domain. Applying the Fourier transform allows the model to capture and balance low- and high-frequency components better. This strategy mitigates spectral bias, and improves the robustness of depth estimation under partial occlusion by preserving critical structural cues.

Figure 4 depicts the proposed FDP, which estimates the depth of objects using the Fourier transform inspired by Chi et al. [41]. The FDP receives integrated multiscale features extracted from three layers of the ResNet-50 backbone. The integrated feature map $X_t^{l,g} \in \mathbb{R}^{C\times H\times W}$ is split along the channel dimension: the local focus feature X_t^l and global focus feature X_t^g, where $X_t^l, X_t^g \in \mathbb{R}^{\frac{C}{2}\times H\times W}$. The local branch preserves fine-grained spatial details using a standard convolution, whereas the global branch captures long-range dependencies in the frequency domain via spectral transformation. The input $X_t^{l,g}$ passes through a fast Fourier convolution, as follows:

$$X_{t+1}^{l,g} = [X_{t+1}^l; X_{t+1}^g] \in \mathbb{R}^{C\times H\times W}, \tag{11}$$

$$X_{t+1}^g = W_1^l * X_t^l + Y_t^g, \tag{12}$$

$$Y_t^g = W_2^g * \left(\mathcal{F}^{-1}\left(W_3^g * \mathcal{F}(W_4^g * X_t^g)\right) + W_4^g * X_t^g \right), \tag{13}$$

$$X_{t+1}^l = W_5^l * X_t^g + W_6^l * X_t^l, \tag{14}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the Fourier and inverse Fourier transforms, respectively, and $*$ represents a convolutional operator. Equation (11) is applied twice in succession to enhance the feature representation, then the resulting output feature is added to the input $X_t^{l,g}$ via a residual connection. Next, this composite block is repeated three times to estimate the depth features progressively.

D. 3D Detection Decoder with Keypoint Embedding

The 3D detection decoder refines queries by sequential cross-attention that integrates depth and keypoint cues. Learnable object queries are initialized from the last-layer outputs of a 2D detection transformer. Visible 2D keypoints $\hat{k}$ are embedded into geometry-aware positional encodings, while depth features are extracted from an FDP. At each decoder layer, the queries first attend to depth features via depth cross-attention, and then to the 2D keypoints $\hat{k}$ embedded into spatial positional encodings via keypoint cross-attention. Inspired by MonoDGP [3], the updated queries undergo self-attention and subsequently interact with multi-scale visual features through visual cross-attention. Finally, the updated queries are processed using a feedforward network and MLP-based detection heads to predict the 3D object boxes B'_{3D}.

E. Relational Graph-based Refinement

The B'_{3D} produced by the 3D detection decoder is predicted independently for each object, which can cause incomplete geometry under occlusion. This work proposes a relational graph refinement module that uses graph reasoning to transfer the higher accuracy of non-occluded objects, refining coarse 3D bounding boxes of occluded ones.

As Fig. 5 illustrates, the relational graph refinement module receives the initial bounding boxes B'_{3D} as input. The refinement process merges predicted bounding boxes based on the degree of overlap. The refined bounding boxes are concatenated with the corresponding $\hat{k}$ and θ obtained from the occlusion-robust 2D keypoint detection module. In the

graph construction block, the concatenated features pass through two linear layers to produce node embeddings N. A pairwise distance matrix is computed based on the 3D bounding box centers, and edges E are formed by connecting pairs of nodes whose centers are adjacent in Euclidean space. The constructed graph features are processed using multiple graph interaction blocks, where each bounding box (node) is influenced by its adjacent bounding box (neighbor nodes) through the E defined in the graph. Each block contains two graph convolutional layers, which aggregate neighborhood features, followed by a linear layer that predicts residual updates ΔB_{3D} for each node. The final refined 3D boxes B_{3D} are obtained by adding these ΔB_{3D} to the B'_{3D}. All 3D detection modules in MonoKey, except for the occlusion-robust 2D keypoint detection, are optimized with a unified 3D detection loss [3].

IV. EXPERIMENTS

A. Datasets

This work presents experiments conducted on the KITTI benchmark [42], comprising 7,481 annotated images. The dataset is split into 3,712 frames for training and 3,769 frames for testing. Following the standard protocol, we evaluated the accuracy using the 3D average precision metrics: AP_{3D} for 3D object detection and AP_{BEV} for the bird's eye view localization, both computed across 40 recall thresholds. The evaluation is performed at an intersection-over-union (IoU) threshold of 0.7.

This work additionally employs the Dense dataset [43], which targets diverse weather conditions such as clear, snowy, rainy, and foggy, and comprises about 10,979 images and it was reformatted to the KITTI style. This work also uses the CADC dataset [44], focused on snowy weather, contains about 7,001 images. Both datasets were evaluated using the AP_{BEV} metric at an IoU threshold of 0.5.

B. Implementation Details

MonoKey was trained for 95 epochs and was configured with a learning rate initialized to 2e-4 and a batch size of 4 on an RTX-4080 GPU. The AdamW optimizer with weight decay was employed. In the occlusion-robust 2D keypoint detection module, the 2D detector and 2D keypoint detector adopted DLA-34 as the feature backbone. Moreover, a symmetry-aware regularization hyperparameter λ_{sym} was introduced and set to 1 during training.

C. Results

We evaluated the performance of the proposed method against SOTA monocular 3D object detection methods via quantitative comparisons and qualitative visualizations to demonstrate its effectiveness. In each table, the bold and underlined fonts indicate the best and second-best scores.

Table I compares the performance of monocular 3D object detectors on the KITTI dataset. While the overall performance improvement over baselines such as MonoDGP appears modest, the gains are more pronounced under moderate and hard levels. Moreover, occlusion is defined by calculating the IoU overlap between 2D bounding boxes, where objects with an IoU overlap ratio of 20%–60% are regarded as occluded cases. The IoU range is set to 20%–60% since it corresponds to the most frequent occlusion ratios in real driving datasets. Under this setting, MonoKey achieves the best performance. This indicates that MonoKey is particularly effective in occlusion scenarios where structural cues are severely degraded, highlighting its practical robustness in real-world autonomous driving environments.

TABLE I: Results of monocular 3D object detectors on the KITTI dataset for the car class.

Metric	$AP_{3D}(\%)\uparrow$				$AP_{BEV}(\%)\uparrow$			
Difficulty	Easy	Moderate	Hard	Occlusion	Easy	Moderate	Hard	Occlusion
DEVIANT [12]	24.63	16.54	14.52	15.35	32.60	23.04	19.99	21.23
MonoDTR [4]	24.52	18.57	15.51	16.12	33.33	25.35	21.68	22.37
MonoDETR [2]	28.84	20.61	16.38	17.64	37.86	26.95	22.80	24.17
MonoCD [11]	26.45	19.37	16.38	17.20	34.60	24.96	21.51	23.02
MonoDGP [3]	<u>30.76</u>	<u>22.34</u>	<u>19.02</u>	<u>20.33</u>	<u>39.40</u>	<u>28.20</u>	<u>24.42</u>	<u>26.54</u>
Ours	**31.61**	**23.39**	**20.19**	**23.38**	**39.78**	**29.42**	**25.81**	**29.29**

TABLE II: Results of monocular 3D object detectors on the CADC and Dense for the car class.

Metric	$AP_{BEV}(\%)$				
Dataset	CADC	Dense			
Weather	Snow	Clear	Rain	Snow	Fog
MonoDETR	<u>2.71</u>	17.58	<u>0.89</u>	<u>1.98</u>	<u>0.76</u>
MonoCD	1.28	16.33	0.87	1.72	0.38
MonoDGP	2.27	<u>18.70</u>	0.25	1.81	0.34
Ours	**9.09**	**18.92**	**1.12**	**4.55**	**3.03**

To further validate robustness under diverse weather, which can also be regarded as a form of partial occlusion, Table II presents the cross-dataset evaluation results, where the models were trained and evaluated on both the CADC and Dense datasets. Compared with recent approaches, our method achieves the highest performance. Although prior methods often collapse to near-zero under severe weather, our approach remains more robust.

To better illustrate the quantitative findings, Figure 6 presents visual comparisons of MonoDGP and MonoKey on the KITTI and CADC datasets, including both image-based results and their BEV representations. While MonoDGP misses predictions for occluded objects, MonoKey demonstrates effective detection under occlusion. In addition, Figure 7 presents the reconstructed keypoints for the car class in the KITTI dataset. In the top row, the three images on the left illustrate cases where keypoints are directly predicted from partially visible input. In contrast, the bottom row shows that the proposed prior-guided autoencoder reconstructs coherent keypoints, restoring the correct object shape even under challenging conditions. The two images on the right depict cases where the object is partially cropped by the image boundary. In the top row, direct predictions fail to infer missing parts, whereas in the bottom row, the proposed method uses structural constraints to recover keypoints close to the original shape, showing robust reconstruction under boundary-induced limitations.

Complementing Fig. 7, Table III reports the keypoint detection accuracy on CarFusion in terms of percentage of correct keypoints (PCK) at thresholds 0.05 and 0.1. Since the KITTI dataset does not provide GT keypoints, the quan-

Fig. 6: Results of the 3D object detectors, (a) MonoDGP, (b) MonoKey, on KITTI and CADC. Purple boxes: missed predictions; blue boxes: correct predictions.

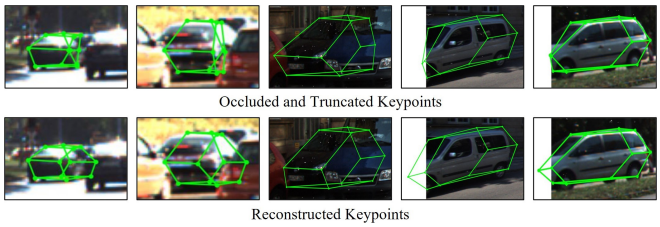


Fig. 7: Visual results of the reconstructed keypoints using the proposed prior-guided autoencoder on KITTI.

TABLE III: Keypoint detection accuracy on CarFusion.

Method	All cases		Occluded cases	
	PCK@0.05 ↑	PCK@0.1 ↑	PCK@0.05 ↑	PCK@0.1 ↑
CenterNet	0.818	0.908	0.719	0.853
Ours	0.879	0.929	0.816	0.896

TABLE IV: MAE of the 3D yaw estimator on KITTI.

Metric	MAE (rad) ↓
MonoDGP	0.270
Ours	0.185

titative evaluation is conducted on CarFusion. Our method achieves higher accuracy than CenterNet in both cases, due to the reconstruction of occluded keypoints. Subsequently, Table IV presents the mean absolute error (MAE) of the 3D yaw estimation. Our method demonstrates more accurate yaw prediction than the baseline.

Table V compares the complexity of monocular 3D object detectors, evaluated in terms of the number of parameters, FLOPs, and latency. Although the integration of keypoint and graph refinement modules slightly increases the number of parameters and computational cost, resulting in a modest overhead in inference time, this additional cost is minor relative to the substantial performance gains under occlusion. To further mitigate latency, our design processes the visual encoder and depth predictor in parallel, while the 2D detection decoder is shared across both the 2D keypoint detector and the 3D detection decoder. With 48M parameters, the model remains within a range that is feasible for deployment on modern embedded platforms (e.g., NVIDIA Jetson Xavier). It can be considered acceptable for real-time robotics applications, where robustness against occlusion is important. Our future work will focus on reducing complexity of the 3D detector backbone for lightweight deployment.

TABLE V: Computational complexity comparison on the KITTI dataset.

Metric	Params (M) ↓	FLOPs (G) ↓	Latency (ms) ↓
MonoDTR	54.27	117.96	62.81
MonoDETR	35.93	119.44	38.87
MonoCD	41.96	171.20	42.36
MonoDGP	38.90	68.99	33.16
Ours	47.84	85.85	39.34

TABLE VI: Ablation study for the MonoKey on KITTI.

Occlusion-robust 2D Keypoint Detection		Frequency-based Global-local Depth Predictor	Relational Graph Refinement	AP_{3D} (%) ↑ (occluded cases)
Raw keypoint	Reconstructed keypoint			
	✓	✓	✓	23.38
✓		✓	✓	22.08
	✓		✓	22.76
	✓	✓		22.89
				20.33

TABLE VII: Detection variation with λ_{sym} on KITTI at moderate difficulty.

λ_{sym}	0	0.5	1	1.5	2
$AP_{3D}\%$ ↑	22.56	22.93	23.39	22.99	22.84

D. Ablation Study

Table VI presents the ablation study that evaluates the effectiveness of each proposed component under occluded cases. In each row, the activated modules are indicated by a check mark (✓). The first row shows the performance when all modules are enabled, achieving the best result. The second, third, and fourth rows correspond to the removal of the reconstructed keypoint, FDP, and relational graph refinement, each leading to a drop in accuracy. The final row represents the baselines, resulting in the lowest performance. The ablation study demonstrates that each module contributes to robustness against occlusion, and further shows that the reconstruction process can realign keypoints correctly even when they are initially misdetected.

Table VII presents the variations in 3D detection performance on the KITTI dataset as the weighting factor λ_{sym} in Eq. (2) is adjusted, highlighting its role in balancing reconstruction fidelity and symmetry constraints. The best performance is achieved when λ_{sym} is set to 1.

V. CONCLUSION

This paper addresses the limitations of monocular 3D object detection under occlusion by proposing a keypoint-based approach, MonoKey, which estimates keypoints and reconstructs occluded regions using prior structural symmetry characteristics. Moreover, MonoKey enhances the depth estimation by incorporating global contextual 3D cues. Further, MonoKey analyzes structural relationships between objects to refine the bounding boxes. Thus, the proposed method handles occlusion, outperforming SOTA monocular 3D object detection approaches. While our work focuses on rigid objects such as vehicles, the framework can naturally extend to non-rigid categories (e.g., cyclists), for which semantic keypoints are currently unavailable.

REFERENCES

[1] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, "Resolution-robust large mask inpainting with fourier convolutions," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 2149–2159.

[2] R. Zhang, H. Qiu, T. Wang, Z. Guo, Z. Cui, Y. Qiao, H. Li, and P. Gao, "Monodetr: Depth-guided transformer for monocular 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 9155–9166.

[3] F. Pu, Y. Wang, J. Deng, and W. Yang, "Monodgp: Monocular 3d object detection with decoupled-query and geometry-error priors," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 6520–6530.

[4] K.-C. Huang, T.-H. Wu, H.-T. Su, and W. H. Hsu, "Monodtr: Monocular 3d object detection with depth-aware transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4012–4021.

[5] D. Park, J. Li, D. Chen, V. Guizilini, and A. Gaidon, "Depth is all you need for monocular 3d detection," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 7024–7031.

[6] Q. Yang, H. Chen, Z. Chen, and J. Su, "Uncertainty estimation for monocular 3d object detectors in autonomous driving," in *ICRAE*. IEEE, 2021, pp. 55–59.

[7] Y. Liu, Z. Xu, and M. Liu, "Star-convolution for image-based 3d object detection," in *ICRA*. IEEE, 2022, pp. 5018–5024.

[8] L. Jing, R. Yu, H. Kretzschmar, K. Li, C. R. Qi, H. Zhao, A. Ayvaci, X. Chen, D. Cower, Y. Li, *et al.*, "Depth estimation matters most: Improving per-object depth estimation for monocular 3d detection and tracking," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 366–373.

[9] C. Picron, P. Chakravarty, T. Roussel, and T. Tuytelaars, "What my motion tells me about your pose: A self-supervised monocular 3d vehicle detector," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 293–13 300.

[10] Y. Chen, L. Tai, K. Sun, and M. Li, "Monopair: Monocular 3d object detection using pairwise spatial relationships," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 093–12 102.

[11] L. Yan, P. Yan, S. Xiong, X. Xiang, and Y. Tan, "Monocd: Monocular 3d object detection with complementary depths," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10 248–10 257.

[12] A. Kumar, G. Brazil, E. Corona, A. Parchami, and X. Liu, "Deviant: Depth equivariant network for monocular 3d object detection," in *European Conference on Computer Vision*. Springer, 2022, pp. 664–683.

[13] L. Wang, L. Du, X. Ye, Y. Fu, G. Guo, X. Xue, J. Feng, and L. Zhang, "Depth-conditioned dynamic message propagation for monocular 3d object detection," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2021, pp. 454–463.

[14] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2019, pp. 8445–8453.

[15] Y. Lei, X. Li, Z. Jiang, X. Ju, and J. Liu, "Aeam3d: Adverse environment-adaptive monocular 3d object detection via feature extraction regularization," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 4135–4139.

[16] X. Li, J. Liu, Y. Lei, L. Ma, X. Fan, and R. Liu, "Monotdp: Twin depth perception for monocular 3d object detection in adverse scenes," *arXiv preprint arXiv:2305.10974*, 2023.

[17] T. Wang, X. Zhu, J. Pang, and D. Lin, "Fcos3d: Fully convolutional one-stage monocular 3d object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 913–922.

[18] H. Chen, Y. Huang, W. Tian, Z. Gao, and L. Xiong, "Monorun: Monocular 3d object detection by reconstruction and uncertainty propagation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10 379–10 388.

[19] 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*, 2021, pp. 8555–8564.

[20] X. Liu, N. Xue, and T. Wu, "Learning auxiliary monocular contexts helps monocular 3d object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 1810–1818.

[21] I. Lee, E. Lee, and S. B. Yoo, "Latent-ofer: Detect, mask, and reconstruct with latent vectors for occluded facial expression recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1536–1546.

[22] P. Šebek, Š. Pokorný, P. Vacek, and T. Svoboda, "Real3d-aug: Point cloud augmentation by placing real objects with occlusion handling for 3d detection and segmentation," *arXiv preprint arXiv:2206.07634*, 2022.

[23] G. Brazil and X. Liu, "M3d-rpn: Monocular 3d region proposal network for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9287–9296.

[24] X. Jiang, S. Jin, X. Zhang, L. Shao, and S. Lu, "Monomae: Enhancing monocular 3d detection through depth-aware masked autoencoders," in *Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 11 392–11 411.

[25] N. D. Reddy, M. Vo, and S. G. Narasimhan, "Carfusion: Combining point tracking and part detection for dynamic 3d reconstruction of vehicles," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1906–1915.

[26] H. Pahadia, D. Lu, B. Chakravarthi, and Y. Yang, "Skope3d: A synthetic dataset for vehicle keypoint perception in 3d from traffic monitoring cameras," in *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2023, pp. 4367–4372.

[27] X. Han, S. Wang, X. Huang, and Z. Kan, "Posefusion: Multi-scale keypoint correspondence for monocular camera-to-robot pose estimation in robotic manipulation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 795–801.

[28] S. W. Hyder, M. Usama, A. Zafar, M. Naufil, F. J. Fateh, A. Konin, M. Z. Zia, and Q.-H. Tran, "Action segmentation using 2d skeleton heatmaps and multi-modality fusion," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 1048–1055.

[29] F. Han, X. Yang, C. Reardon, Y. Zhang, and H. Zhang, "Simultaneous feature and body-part learning for real-time robot awareness of human behaviors," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 2621–2628.

[30] C. Zimmermann, T. Welschehold, C. Dornhege, W. Burgard, and T. Brox, "3d human pose estimation in rgbd images for robotic task learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1986–1992.

[31] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.

[32] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 1145–1153.

[33] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, *et al.*, "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.

[34] V. Somers, A. Alahi, and C. D. Vleeschouwer, "Keypoint promptable re-identification," in *European Conference on Computer Vision*. Springer, 2024, pp. 216–233.

[35] R. Hachiuma, F. Sato, and T. Sekii, "Unified keypoint-based action recognition framework via structured keypoint pooling," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 962–22 971.

[36] T. Jakab, R. Tucker, A. Makadia, J. Wu, N. Snavely, and A. Kanazawa, "Keypointdeformer: Unsupervised 3d keypoint discovery for shape control," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 783–12 792.

[37] Y. Zhou, Y. He, H. Zhu, C. Wang, H. Li, and Q. Jiang, "Monoef: Extrinsic parameter free monocular 3d object detection," *IEEE Trans-*

actions on Pattern Analysis and Machine Intelligence, vol. 44, no. 12, pp. 10 114–10 128, 2021.

[38] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6569–6578.

[39] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2403–2412.

[40] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, "A kernel method for the two-sample-problem," *Advances in neural information processing systems*, vol. 19, 2006.

[41] L. Chi, B. Jiang, and Y. Mu, "Fast fourier convolution," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4479–4488, 2020.

[42] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The international journal of robotics research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[43] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather," in *CVPR*, 2020, pp. 11 682–11 692.

[44] M. Pitropov, D. E. Garcia, J. Rebello, M. Smart, C. Wang, K. Czarnecki, and S. Waslander, "Canadian adverse driving conditions dataset," *The International Journal of Robotics Research*, vol. 40, no. 4-5, pp. 681–690, 2021.