ITC 2/55
Information Technology and Control
Vol. 55 / No. 2/ 2026
pp. 770-783
DOI 10.5755/j01.itc.55.2.43442

Monocular 3D Object Detection Using Multi-Scale Context Fusion Attention

Received 2025/10/27 | Accepted after revision 2026/02/26

HOW TO CITE: Yu, S. (2026) Monocular 3D Object Detection Using Multi-Scale Context Fusion Attention. *Information Technology and Control*, 55(2), 770-783. https://doi.org/10.5755/j01.itc.55.2.43442

Monocular 3D Object Detection Using Multi-Scale Context Fusion Attention

Sixu Yu
School of Media and Communication, Shanghai Jiao Tong University, Shanghai, 200240, China

Corresponding author: yusixu199312@sjtu.edu.cn

Monocular 3D object detection aims to infer three-dimensional spatial information from a single RGB image, but existing approaches are often challenged by depth ambiguity and insufficient interaction between global and local features. To mitigate these issues, a monocular 3D object detection framework based on multi-scale context fusion attention is proposed in this paper. Multi-scale features are first extracted to capture objects at different spatial resolutions. Subsequently, global self-attention is employed to model long-range contextual dependencies, while a local context fusion module aggregates neighborhood features based on similarity-aware interactions, enabling more discriminative local representations. Through the coordinated global–local attention mechanism, depth reasoning capability is enhanced and background interference is effectively suppressed. Experiments conducted on the KITTI benchmark demonstrate that the proposed method achieves 27.12 / 17.68 / 14.21 AP for 3D detection and 34.72 / 23.58 / 19.67 AP for BEV detection, outperforming representative monocular baselines such as MonoDETR. These results indicate that the proposed approach improves detection accuracy and robustness, particularly for distant and partially occluded objects.

KEYWORDS: Monocular 3D Object Detection; Multi-Scale Feature Fusion; Attention Mechanism; Context Modeling; Autonomous Driving

1. Introduction

Three-dimensional object detection has become a critical foundation for autonomous driving, medical imaging, and intelligent robotics. However, current methods still struggle with issues such as sparse depth cues, occlusion, lighting variation, and cross-domain adaptability. A large number of recent studies have attempted to address these challenges from different sensing, modeling, and application perspectives.

Nurfauzi et al. [26] focused on the automatic detection of traumatic bleeding in whole-body CT, introducing a 3D object detection model to improve

lesion localization accuracy and accelerate radiological diagnosis. Noh et al. [25] proposed 3DPillars, a pillar-based two-stage 3D detection framework that enhances localization stability under occlusion and sparse depth information. In the context of mining environments, Essien and Frimpong [11] reviewed 3D object detection systems for underground autonomous trucks, summarizing technical bottlenecks in low-light and dusty tunnels. Dhakal et al. [6] introduced VirtualPainting, a data-augmentation approach using virtual points and distance-aware sampling to mitigate sparsity and improve far-range detection. Boyko [4] developed an efficient RGB-D–based method for object detection and localization in autonomous robotic systems, achieving both real-time performance and high spatial precision.

Beyond visible light sensing, Jaber et al. [18] presented 3D-DUO, a detection network for low-resolution multibeam echosounder data, enabling robust underwater object localization. Yildiz et al. [17] fused camera and LiDAR data through semantic depth sensing, achieving accurate distance estimation and improved real-time detection. Ramana et al. [29] enhanced multi-scale recognition via Spectral Pyramid Pooling and fused keypoint features embedded in a ResNet-50 backbone, strengthening feature discrimination for complex objects. Zakeri et al. [37] constructed SMS3D, a synthetic 3D mushroom-scene dataset for joint detection and pose estimation, filling an important gap in domain-specific benchmarks. Kozov et al. [20] comparatively analyzed display technologies for 3D defect visualization, providing guidelines for selecting effective visualization hardware.

Fawole and Rawat [12] comprehensively surveyed 3D object detection methods for self-driving vehicles, highlighting the rapid evolution from handcrafted geometry to deep feature fusion. Soans and Fukumizu [33] designed an anchor-free detection framework for synthetic 3D traffic-sign datasets, integrating depth estimation and character recognition for improved detection robustness. Park et al. [21] proposed a motion-aware temporal fusion network to capture dynamic LiDAR sequences, improving detection stability for moving targets. B. A N M et al. [3] quantified the effect of LiDAR point-cloud compression on detection accuracy, offering practical trade-offs between data size and precision. Shi et al. [27] introduced CenRadFusion, combining image-center detection and millimeter-wave radar for cross-modal fusion, effectively enhancing detection under challenging conditions.

Several studies have refined network architectures to handle sparsity and efficiency. Erabati and Araujo [10] proposed SRFDet3D, a sparse-region fusion approach that strengthens feature representation in difficult areas, while Naich and Carrión [24] developed an intensity-aware LiDAR detection model to exploit reflectance variations. Erabati and Araujo [9] further designed DeLiVoTr, a lightweight voxel transformer reducing computational cost without degrading accuracy. Zamanakos et al. [28] introduced Feature-Aware Re-weighting (FAR) in bird's-eye-view representations, adaptively emphasizing informative spatial regions. Contreras et al. [5] conducted a real-time survey of 3D object detection methods for autonomous driving, summarizing latency–accuracy trade-offs and design principles.

In industrial automation, Reuse et al. [31] explored factory-driving scenarios with limited annotations and proposed ground-truth sampling augmentation to mitigate data scarcity. Tiwari and Sharma [34] presented FS-3DSSN, an efficient few-shot single-stage detector for point clouds, ensuring stable performance with minimal samples. Joshi et al. [19] applied 3D integral imaging and deep learning to underwater object and temporal-signal detection in turbid water, maintaining reliable recognition under severe scattering. Alaba et al. [1] reviewed multimodal fusion trends for autonomous-vehicle perception, analyzing sensor-fusion architectures for 3D object detection.

Multi-sensor and vision-based comparisons continue to expand. Housam et al. [14] evaluated video–LiDAR fusion through ablation studies, showing complementary temporal–spatial benefits. Alexandre et al. [2] improved 3D monocular detection and tracking efficiency for smart mobility using optimized transformer fusion. Uwe and Yusheng [36] investigated urban-object change detection using 3D point clouds, emphasizing temporal alignment accuracy. Ghazal et al. [13] proposed a deep visual-attention model for saliency detection on 3D objects, reinforcing fine-detail understanding. Duy and Linh [8] generated pseudo-LiDAR via simple linear iterative clustering for low-cost 3D detection, suitable for budget-con-

strained systems. Finally, Kashif et al. [35] compared passive integral imaging and active LiDAR sensing under fog and occlusion, demonstrating complementary resilience of hybrid systems.

Collectively, these studies demonstrate steady progress in 3D object detection across sensing modalities, environments, and architectures. Nevertheless, three major challenges remain for monocular 3D detection:

1 Limited global reasoning – many models rely on local receptive fields, lacking mechanisms to capture long-range spatial dependencies;
2 Weak local detail preservation – fine-grained cues near boundaries or small objects are often overshadowed by global aggregation; and
3 Monocular depth ambiguity – without depth sensors, robust geometric inference and cross-scale fusion become essential.

The main contributions of this work are summarized as follows:

1 A global–local context fusion attention framework is proposed for monocular 3D object detection, in which global self-attention and local context fusion are jointly modeled to enhance spatial reasoning across multiple scales.
2 A similarity-aware local context fusion strategy is introduced, leveraging clustering-based neighborhood aggregation and cosine-similarity attention to preserve fine-grained object details while suppressing background interference.
3 Extensive experiments on the KITTI benchmark demonstrate that the proposed method consistently outperforms representative monocular baselines, particularly in challenging scenarios involving distant and partially occluded objects.

2. Theoretical Background

2.1. Design and Implementation of the Global Self-Attention Module

In the global self-attention module, image patch division and tokenization are the key steps. The main idea is to convert the original input image into a sequence of fixed-dimensional tokens, making it suitable for continuous attention-based processing. Let the image size be $H\times W$ pixels. For convenience, the image is divided into small patches of size $P\times P$. Hence, the total number of patches N can be calculated as:

$$N=\frac{H\times W}{P^2}. \tag{1}$$

Each patch P_i contains $P\times P$ pixels, where C denotes the number of image channels. We define the patch set as $\{P_i\}_{i=1}^{N}$, where $P_i\in R^{h\times w\times c}$ represents the pixel features of the i-th patch.For each center patch P_c, this study defines its 8 neighboring patches as the neighborhood set $N(P_c)$. The feature representation of the neighborhood center is expressed as:

$$F_{center}=\frac{1}{|N(P_c)|}\sum_{P_j\in N(P_c)}F(P_j). \tag{2}$$

To transform each patch into a token, the patch is first flattened into a one-dimensional vector. Then, a linear projection is applied to map it into a higher-dimensional token space through the learnable weight matrix W_p and bias term b_p:

$$t_i=W_p\cdot vec(P_i)+b_p,\ \ i=1,2,\ldots,N. \tag{3}$$

Here, $vec(P_i)$ denotes the flattening operation of patch P_i, and $t_i\in R^D$ is the generated token with dimension D. Thus, the original image, after patch division and tokenization, is converted into a token sequence:

$$T=[t_1,t_2,\ldots,t_N]\in R^{N\times D}. \tag{4}$$

To make the model more sensitive to spatial position information, positional encoding $E_{pos}\in R^{N\times D}$ is added. It can be generated either in a fixed or learnable manner. The final input is obtained by summing the token sequence and its positional encoding:

$$X=T+E_{pos}. \tag{5}$$

This positional encoding enables the model to distinguish between different relative positions of patches during attention computation, thereby capturing the global spatial structure. The fixed sinusoidal positional encoding is expressed as:

$$E_{pos}(i,2j)=sin\left(\frac{i}{10000^{\frac{2j}{d}}}\right) \quad (6)$$

$$E_{pos}(i,2j+1)=cos\left(\frac{i}{10000^{\frac{2j}{d}}}\right), \quad (7)$$

where i denotes the token index and j represents the dimension index. The positional encoding can also be defined as a learnable parameter and updated through backpropagation to adapt to specific tasks.After tokenization, the global feature aggregation module performs information interaction among all tokens via the attention mechanism, enabling aggregation and capture of global representations.According to the standard query–key–value structure, multi-head self-attention captures diverse global dependencies in the feature space:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (8)$$

Each attention head uses learnable projection matrices $W_i^Q, W_i^K, W_i^V \in R^{N\times d_k}$ for the i-th head, where $d_k=\frac{D}{h}$. The output of each head is computed as:

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V). \quad (9)$$

The outputs of all heads are concatenated to form the global feature representation:

$$Z= Concat(head_1,...,head_h)\in R^{N\times D}. \quad (10)$$

To better capture nonlinear high-level features, the output of the multi-head attention layer is passed through a Feed-Forward Network (FFN), which typically contains two linear transformations and one activation function (ReLU):

$$Z'=ReLU(ZW_1+b_1)W_2+b_2, \quad (11)$$

where $W_1\in R^{D\times d_{ff}}$, $W_2\in R^{d_{ff}\times D}$ are weight matrices, b_1, b_2 are bias terms, and d_{ff} is the hidden layer dimension. This design allows the model to further extract deep semantic features through nonlinear transformations, enhancing its expressive capability.

To prevent overfitting and stabilize training, residual connections and layer normalization are applied after each multi-head attention and FFN sublayer. Given the input X, the outputs after residual connection and normalization are:

$$Y=LayerNorm(X+Z) \quad (12)$$

$$O=LayerNorm(Y+Z'). \quad (13)$$

Through this mechanism, the inter-token dependencies and interactions are effectively represented via attention weight matrices.

2.2. Local Context Fusion and Adaptive Weight Allocation Strategy

2.2.1. Local Neighborhood Feature Extraction and Information Aggregation

In the local context fusion module, to better capture fine-grained information within the target region, this study adopts a local neighborhood feature extraction and information aggregation approach to efficiently collect neighborhood information.

Let the input image be denoted as $I\in R^{H\times W\times C}$, where H and W represent the image height and width, and C is the number of channels.

The feature extraction target aims to enhance and reallocate spatial attention to form a feature map $F_w\in R^{H'\times W'\times C'}$, enabling the model to better aggregate fine-grained salient regions in the local context.

First, the input image I is divided into P×Psmall patches.

The total number of divided patches is $N=\frac{H}{h}\times\frac{W}{w}$. Each patch $P_i\in R^{h\times w\times c}$ represents pixel features of the i-th region. For each patch center P_c, this study defines its 8 adjacent patches as a neighborhood set $N(P_c)$.Neighborhood feature is computed following Equation (2).

Then, the patch features are further clustered using the K-means algorithm to minimize intra-cluster Euclidean distance.

The cluster centers are denoted as $\{C_k\}_{k=1}^K$, where each $C_k\in R^c$ represents the center of cluster k.The weight of each patch relative to the cluster center is calculated as:

$$w_i=\frac{1}{K}\sum_{k=1}^{K} exp\left(\frac{-\|F_{center,i}-C_k\|^2}{2\sigma^2}\right), \quad (14)$$

where $F_{center,i}$ denotes the feature of the i-th patch center, and σ is the standard deviation parameter controlling sensitivity to feature variation.The weight w_i reflects the saliency of the patch with respect to different cluster centers — the higher the value, the more important the patch is to the local region.

To further strengthen the expressive power of salient regions, each patch feature is multiplied by its adaptive weight:

$$F_{w,i} = w_i \cdot F(P_i). \quad (15)$$

All weighted features are then reassembled to form the feature map F_w, consistent in dimension with the original input feature map. The enhanced feature map F_w contains both local salient information and global contextual background cues. To validate the method's effectiveness, F_w is fed into the detection head of Sparse R-CNN for object detection.

To enhance performance, a saliency enhancement loss is introduced into the total detection loss function as:

$$L_{saliency}=-\frac{1}{N}\sum_{i=1}^{N} \| F_{w,i}-F(P_i) \|^2 . \quad (16)$$

This term penalizes excessive deviation between weighted and original features, ensuring consistency and emphasizing representation of salient areas.

The final detection loss is defined as:

$$L=L_{cls}+\lambda L_{reg}+\lambda L_{saliency} , \quad (17)$$

where L_{cls} and L_{reg} represent classification and regression losses, respectively, and $\lambda L_{saliency}$ is the weighting coefficient for the saliency loss. In summary, the local feature extraction and aggregation module employs adaptive weight allocation and variance-based optimization to capture precise and effective local contextual information.

This approach not only strengthens feature representation within local regions but also fully utilizes multi-layer contextual cues to enhance global–local information fusion, providing richer and more accurate 3D object detection representations.

2.2.2. Local Attention Weight Calculation Based on Cosine Similarity

The local attention weighting based on cosine similarity calculates fine-grained weight distribution by measuring the similarity between the central pixel and other pixels in its neighborhood.

Specifically, let the feature vector of the center pixel be f_c, and the feature vector of any neighboring pixel i be f_i. The cosine similarity between them is expressed as:

$$sim(f_c,f_i)=\frac{f_c \cdot f_i}{\|f_c\|\|f_i\|} . \quad (18)$$

Since cosine similarity values lie within [-1,1], normalization is typically applied. For the neighborhood N(c), this study uses the softmax function to normalize all similarity scores and obtain the final attention weights:

$$\alpha_i=\frac{exp\left(\frac{sim(f_c,f_i)}{\tau}\right)}{\sum_{j\in N(c)} exp\left(\frac{sim(f_c,f_i)}{\tau}\right)} . \quad (19)$$

Here, τ is a temperature parameter controlling the sharpness of the weight distribution – a smaller τ results in sharper, more focused attention.

To ensure numerical stability, each local region's feature vectors are first L2-normalized:

$$\tilde{f}=\frac{f}{\|f\|} . \quad (20)$$

The cosine similarity between normalized vectors is simplified as:

$$sim\ (\tilde{f_c},\tilde{f_i})=\tilde{f_c} \cdot \tilde{f_i} . \quad (21)$$

The obtained attention weights are then used to aggregate local neighborhood information, yielding the local feature representation of the center pixel:

$$\hat{F}(c)=\sum_{i\in N(c)}\alpha_i\cdot f_i\ . \tag{22}$$

This cosine-similarity-based attention mechanism effectively captures fine-grained local correlations, as the weight distribution reflects the angular distance between feature vectors, enhancing local detail perception.

To further strengthen the expressive capability, a learnable linear transformation is applied to the feature vectors before computing cosine similarity:

$$f'=W\cdot f\ . \tag{23}$$

After transformation and re-normalization, the cosine similarity and attention aggregation are recalculated, capturing more discriminative local features.

The final local attention module is integrated into the multi-level feature fusion network to refine localized feature representations, while residual connections ensure that original information is preserved and not completely lost.

2.3. Global–Local Context Fusion Mechanism

To explicitly model the interaction between global and local representations, this study introduces a global–local context fusion mechanism that tightly couples the Global Self-Attention (GSA) module with the Local Context Fusion (LCF) module. Unlike conventional sequential designs in which global and local modules operate independently, the proposed framework establishes a guided fusion strategy, where global contextual information actively modulates local feature aggregation.

Specifically, the GSA module first encodes the input feature maps into a set of token embeddings and captures long-range dependencies through multi-head self-attention. The resulting global representation encodes holistic scene semantics and structural relationships across spatial scales. Rather than being directly concatenated or added to local features, this global contextual embedding is used as a guidance signal for the subsequent local context fusion process.

In the LCF module, local neighborhood features are aggregated through clustering-based adaptive weighting and cosine-similarity attention. During this process, the similarity computation and weight allocation for each local neighborhood are conditioned on the global contextual representation produced by GSA. Concretely, the global embedding influences the similarity-aware weighting of local features by adjusting their relative importance with respect to the overall scene context. This design ensures that local feature refinement is informed by global semantics, enabling salient object regions to be emphasized while suppressing background interference.

Through this global-to-local information flow, the proposed fusion mechanism preserves fine-grained spatial details while maintaining global structural awareness. As a result, the model achieves a balanced representation that integrates long-range contextual reasoning with precise local detail modeling. This coupling strategy fundamentally differs from simple sequential attention pipelines and plays a crucial role in enhancing depth reasoning, robustness under occlusion, and detection accuracy in monocular 3D object detection.

3. Overall Algorithm Framework

The overall architecture of the proposed monocular 3D object detection framework is illustrated in Figure 1. The framework is designed to jointly model global contextual dependencies and local structural details through a multi-scale context fusion attention mechanism.

Given a single RGB image as input, a multi-scale feature extraction backbone is first employed to generate hierarchical feature maps at different spatial resolutions. These feature maps capture objects of varying sizes and provide a foundation for robust depth reasoning.

Subsequently, the extracted multi-scale features are fed into the Global Self-Attention (GSA) module, where long-range dependencies among spatial regions are explicitly modeled. By tokenizing feature maps into patch embeddings and applying multi-head self-attention with positional encoding, the GSA module enables global information exchange across the entire scene, enhancing holistic semantic understanding and improving depth-aware feature representation.

In parallel, the Local Context Fusion (LCF) module focuses on refining fine-grained local details that may be weakened during global aggregation. For each local region, neighborhood features are adaptively aggregated

based on clustering-aware weighting and cosine similarity–based attention. This design allows salient object boundaries and local geometric cues to be preserved while suppressing irrelevant background responses.

Importantly, the proposed framework adopts a global–local cooperative fusion strategy, in which the globally enhanced features from the GSA module guide the local feature refinement process in the LCF module. The outputs of both modules are fused in a complementary manner, ensuring consistency between global scene context and local object structure.

Finally, the fused features are delivered to the 3D detection head, which jointly predicts the object category, 3D location, dimensions, and orientation. The entire framework is trained end-to-end, allowing all modules to be optimized collaboratively.

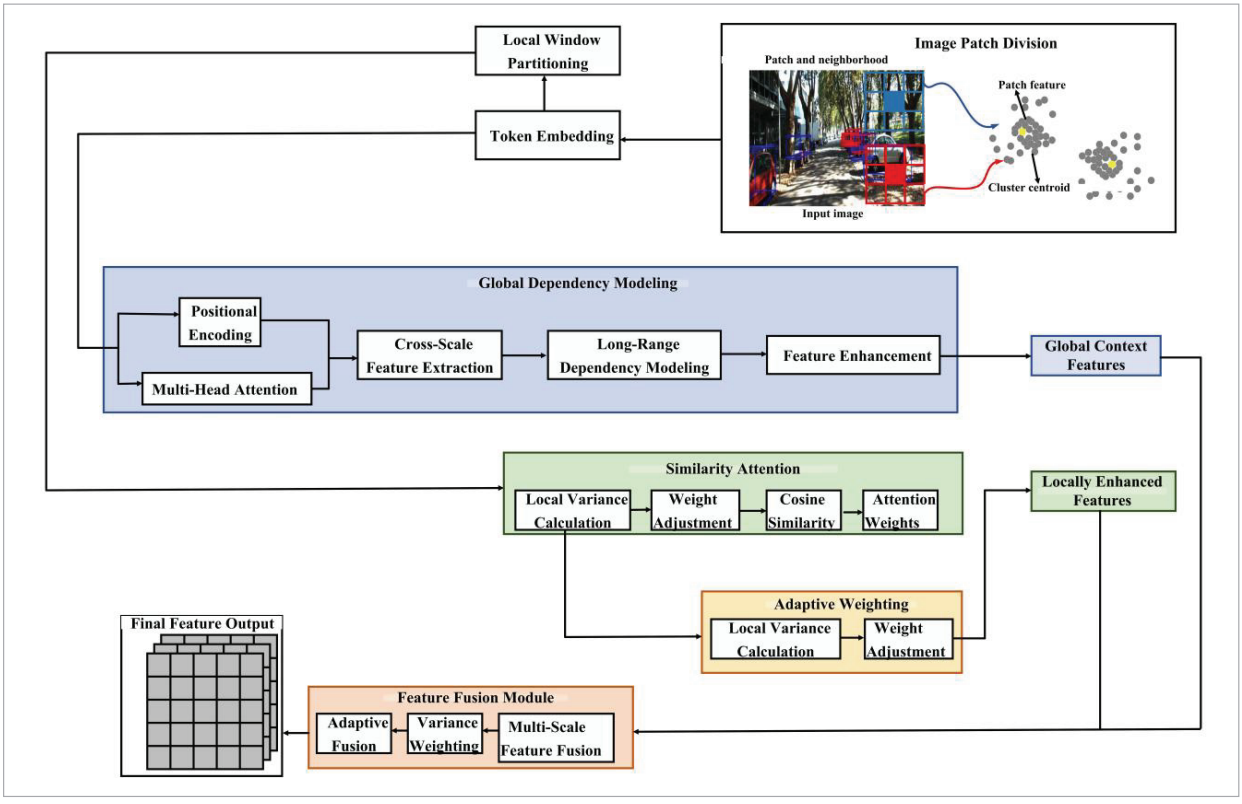
Through this structured global–local interaction, the proposed method effectively balances scene-level reasoning and fine-grained detail modeling, leading to improved accuracy and robustness in monocular 3D object detection.

4. Experimental Results and Analysis

In this section, the proposed method is evaluated on the KITTI benchmark, a widely used dataset for autonomous 3D object detection in driving scenarios. The KITTI dataset provides stereo images and LiDAR scans with precise 3D ground-truth annotations for three object categories: cars, pedestrians, and cyclists. The dataset contains 7,481 training images, 3,769 validation images, and 7,518 test images. Following standard practice, the official training split is used, and results are reported on both the validation and test sets. We present detection results under three difficulty levels – Easy, Moderate, and Hard – and evaluate performance using the Average Precision (AP) metric for both 3D bounding boxes and bird's-eye view (BEV) representations. The proposed model is implemented using PyTorch and trained on a workstation equipped with an NVIDIA A6000 GPU. The training process spans 195 epochs with a batch size of 16 and an initial learning rate of

Figure 1

Framework of the Multi-Scale Context Fusion Attention Mechanism Model.



0.001. The dataset settings and experimental environment configuration are kept consistent across all experiments discussed in Chapters 4-5.

4.1. Comparative Experimental Results and Analysis

To provide a comprehensive comparison, ten representative monocular 3D object detection methods from recent literature were selected for evaluation, each reflecting a different development direction within the field.

Contextual Patch-NetVLAD [15] introduces a contextual feature aggregation strategy based on local patch descriptors and NetVLAD encoding, aiming to enhance spatial correlation modeling; however, its limited depth reasoning capability constrains performance on distant or occluded targets. D4LCN [40] employs a depth-guided convolutional network that couples depth prediction with detection tasks, improving feature localization but still suffering from scale inconsistency. DDMP-3D [7] utilizes dense depth-map prediction with 3D geometric projection to strengthen monocular depth cues, which enhances performance but increases model complexity. MonoDTR [22] incorporates transformer-based feature relation modeling to capture long-range dependencies, achieving improved global context perception in 3D space.

MonoDLE [35] refines feature alignment through depth-level estimation (DLE), enhancing localization consistency across multi-scale feature maps. MonoRCNN [39] extends the region-based detection framework by adding 3D geometric reasoning branches, leading to better regression accuracy in object orientation and size. MonoGeo [30] introduces geometric constraint learning to improve the interpretability of spatial reasoning, effectively handling complex perspective distortions. MonoFlex [21] integrates a flexible depth representation mechanism, achieving balance between geometric precision and computational efficiency.

GUPNet [28] leverages geometry uncertainty propagation, explicitly modeling prediction variance to enhance the reliability of 3D bounding box estimation. Finally, MonoDETR [16] applies a transformer-based end-to-end detection architecture with cross-scale attention, representing the current state-of-the-art baseline in monocular 3D detection. These ten methods collectively illustrate the evolution from handcrafted geometric priors to learning-based glob-

Table 1
The comparison results of different methods on the KITTI dataset.

Methods	Test			Test			Val		
	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
Contextual Patch-NetVLAD **[15]**	15.68	11.12	10.17	22.97	16.86	14.97	-	-	-
D4LCN **[40]**	16.65	11.72	9.51	22.51	16.02	12.55	-	-	-
DDMP-3D **[7]**	19.71	12.78	9.80	28.08	17.89	13.44	-	-	-
MonoDTR **[22]**	21.99	15.39	12.73	28.59	20.38	17.14	24.52	18.57	15.51
MonoDLE **[35]**	17.23	12.26	10.29	24.79	18.89	16.00	17.45	13.66	11.68
MonoRCNN **[39]**	18.36	12.65	10.03	25.48	18.11	14.10	16.61	13.19	10.65
MonoGeo **[30]**	18.85	13.81	11.52	25.86	18.99	16.19	18.45	14.48	12.87
MonoFlex **[21]**	19.94	13.89	12.07	28.23	19.75	16.89	23.64	17.51	14.83
GUPNet **[28]**	20.11	14.20	11.77	22.76	16.46	13.72	-	-	-
MonoDETR **[16]**	25.00	16.47	13.58	33.60	22.11	18.60	28.84	20.61	16.38
The presented model	27.12	17.68	14.21	34.72	23.58	19.67	29.89	21.52	17.25
Improvement value	+2.12	+1.21	+0.63	+1.12	+1.47	+1.07	+1.05	+0.91	+0.87

al–context modeling, showing steady progress in accuracy and robustness. The proposed model in this chapter builds upon this foundation by introducing a multi-scale context fusion attention mechanism, which integrates global self-attention with local feature enhancement to further improve depth reasoning, feature representation, and detection precision.

Table 1 presents the comparison results of different methods on the KITTI dataset, covering 3D detection and bird's-eye view (BEV) detection on the Test set, as well as 3D detection on the Validation set. Each metric is reported under the three official difficulty levels: Easy, Moderate (Mod.), and Hard. Overall, all methods achieve relatively high detection accuracy under the Easy setting, while performance declines progressively as the difficulty increases to Moderate and Hard. Traditional methods such as Contextual Patch-NetVLAD, D4LCN, and DDMP-3D yield only 15.68, 16.65, and 19.71 AP (Easy) on the Test set, and even lower values for Moderate and Hard. Their BEV detection results also fail to show competitive performance.

In contrast, deep learning-based methods such as MonoDTR, MonoDLE, MonoRCNN, MonoGeo, and MonoFlex demonstrate overall improvements, though certain limitations still remain. Among them, MonoDETR achieves 25.00, 16.47, and 13.58 AP for 3D detection on the Test set, and 33.60, 22.11, and 18.60 for BEV detection, while its Validation set scores reach 28.84, 20.61, and 16.38.These results indicate that deep learning-based monocular 3D object detection methods have gradually improved in accuracy over recent years.

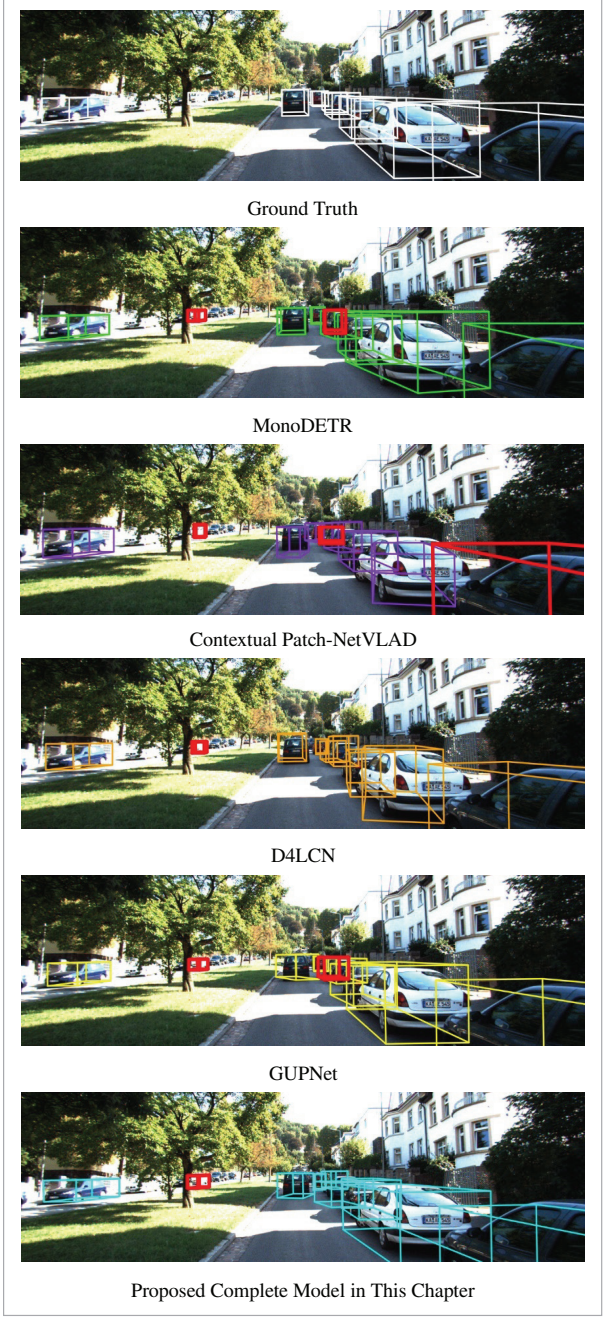
The complete model proposed in this chapter achieves 27.12, 17.68, and 14.21 AP for 3D detection, and 34.72, 23.58, and 19.67 for BEV detection on the Test set, while reaching 29.89, 21.52, and 17.25 on the Validation set — all outperforming the other compared methods.

Specifically, compared with MonoDETR, the proposed model improves the Test set 3D detection results by 2.12, 1.21, and 0.63 percentage points, and improves the Test set BEV results by 1.12, 1.47, and 1.07 percentage points. On the Validation set, the model further achieves 1.05, 0.91, and 0.87 percentage point gains, respectively.

Beyond numerical improvements, this performance gain can be attributed to the fundamental difference in local context modeling between the proposed method and MonoDETR. MonoDETR primarily relies on transformer-based global self-attention and cross-scale feature interaction, in which local information is implicitly encoded within the token-wise

Figure 2

Comparison of detection results across different models. (Red boxes indicate detection errors, including false positives, missed detections, or inaccurate localization.)



attention mechanism. In contrast, the proposed approach introduces an explicit Local Context Fusion (LCF) module, where local neighborhoods are explicitly constructed and refined through clustering-based adaptive weighting and cosine-similarity attention. Moreover, the local aggregation process is guided by global semantic representations obtained from the Global Self-Attention (GSA) module, enabling fine-grained local features to be selectively enhanced under global contextual constraints. This explicit global–local coupling mechanism allows the proposed model to better preserve object boundaries and small-scale structures, thereby improving detection robustness for distant and partially occluded targets.

From the visualization results, it can be observed that the first image represents the ground truth, while the remaining five images show the detection outcomes of different methods under the same scenario. The red boxes highlight areas where detection errors or inaccuracies occur. MonoDETR performs relatively well in detecting nearby vehicles but exhibits unstable recognition for distant targets or those under tree shadows, leading to localization shifts or missed detections. Contextual Patch-NetVLAD shows certain improvements in local detail extraction but still misclassifies vehicles under shaded regions, resulting in more red error boxes.

D4LCN demonstrates noticeable localization deviations for most vehicles, particularly when vehicle edges blend with the background, causing frequent false detections. GUPNet delivers more balanced overall performance; however, in areas with dense trees, it still struggles to distinguish vehicles from the background, leading to multiple red error boxes. In contrast, the proposed complete model produces bounding boxes that align more closely with the actual vehicle positions and maintains relatively accurate localization even for partially occluded or distant targets, with a significantly reduced number of red error boxes.

This indicates that the proposed method demonstrates stronger adaptability in multi-scale feature extraction and local context fusion, effectively distinguishing object contours and positions under complex illumination and background conditions.

Consequently, it substantially reduces both false and missed detections. Overall, this visualization comparison highlights the improved accuracy and robustness of the proposed method, providing more reliable perception capability for autonomous driving applications.

4.2. Ablation Study Results and Parameter Analysis

Table 2 presents the ablation study results of the proposed complete model on the KITTI dataset, focusing on the Global Self-Attention (GSA) module and the Local Context Fusion (LCF) module.

In the experiments, the complete model achieves 3D detection AP on the Test set of 27.12 (Easy), 17.68 (Mod.), and 14.21 (Hard), and BEV detection AP of 34.72 (Easy), 23.58 (Mod.), and 19.67 (Hard). On the Validation set, the corresponding 3D detection AP values are 29.89, 21.52, and 17.25, respectively. By contrast, when the Global Self-Attention module is removed, the Test set results drop to 24.83, 15.41, and 12.52 for 3D detection, 32.31, 22.05, and 18.21 for BEV detection, and the Validation set results decrease to 27.58, 19.96, and 16.12.

When the Local Context Fusion module is removed, the Test set achieves 25.42, 16.21, and 13.35 for 3D detection, 33.68, 22.75, and 19.08 for BEV detection,

Table 2
Ablation study results of the proposed complete model on the KITTI dataset.

Methods	Test AP_{3D}			Test AP_{BEV}			Val AP_{3D}		
	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
Proposed Complete Model	27.12	17.68	14.21	34.72	23.58	19.67	29.89	21.52	17.25
w/o Global Self-Attention Module	24.83	15.41	12.52	32.31	22.05	18.21	27.58	19.96	16.12
w/o Local Context Fusion Module	25.42	16.21	13.35	33.68	22.75	19.08	28.35	20.73	16.82
Improvement over w/o Global SA	+2.29	+2.27	+1.69	+2.41	+1.53	+1.46	+2.31	+1.56	+1.13
Improvement over w/o Local CF	+1.70	+1.47	+0.86	+1.04	+0.83	+0.59	+1.54	+0.79	+0.43

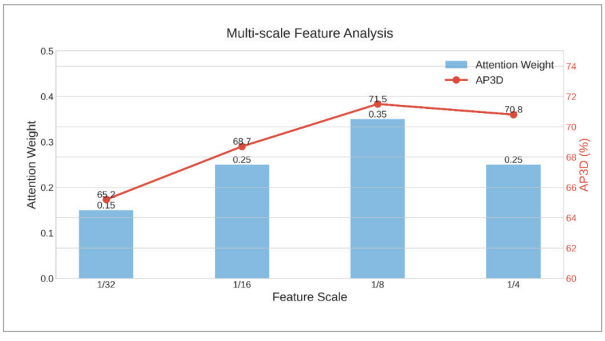
and the Validation set results are 28.35, 20.73, and 16.82, respectively.

From the observed performance gains, compared with the model without the Global Self-Attention module, the complete model improves by 2.29, 2.27, and 1.69 percentage points on Test set 3D detection, by 2.41, 1.53, and 1.46 percentage points on Test set BEV detection, and by 2.31, 1.56, and 1.13 percentage points on the Validation set. Compared with the model without the Local Context Fusion module, the complete model achieves gains of 1.70, 1.47, and 0.86 on Test set 3D detection; 1.04, 0.83, and 0.59 on BEV detection; and 1.54, 0.79, and 0.43 on the Validation set. These results clearly indicate that both modules positively contribute to model performance, but the Global Self-Attention module plays a more decisive role, as its removal leads to more significant performance degradation than that of the Local Context Fusion module. This finding highlights the critical importance of the global attention mechanism in helping the model integrate comprehensive scene-level features, thereby enhancing both 3D object detection and BEV detection accuracy.

In comparison, the Local Context Fusion module mainly contributes to the refinement of fine-grained local features. Although its role is crucial, its impact on overall performance improvement is relatively smaller, as it focuses on optimizing local details and compensating for boundary information. Experimental results further show that, under the Easy difficulty setting, removing either the global self-attention or local fusion module results in a substantial drop in performance – likely because in simpler backgrounds where objects are clearer, precise global–local feature integration becomes particularly vital for accurate localization. Under the Moderate and Hard conditions, where occlusions and background complexity increase, the synergistic effect between the two modules becomes more pronounced; removing either leads to insufficient feature representation, directly reducing detection accuracy.

Figure 3

Multi-scale Feature Analysis.



Overall, compared with its ablated variants, the complete model, which integrates both Global Self-Attention and Local Context Fusion, achieves effective fusion of global scene understanding and local detail representation, delivering superior performance across all evaluation metrics.

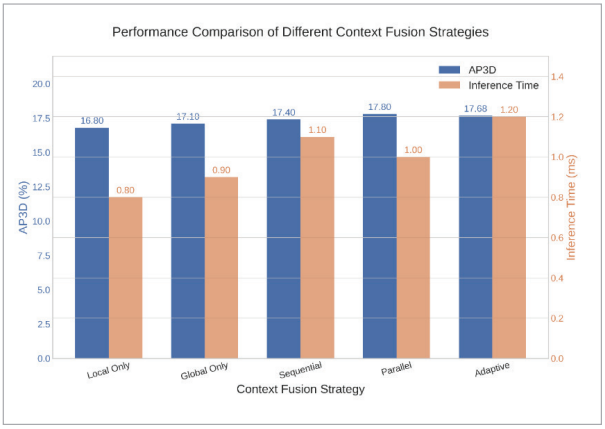
The results further validate the core role of these two modules in enhancing model robustness and generalization capability, and demonstrate that their cooperative interaction significantly improves detection performance across different difficulty levels.

In future research, the fusion strategy between global and local information could be further optimized, and more efficient attention mechanisms may be explored to maintain model lightweightness while achieving higher detection accuracy, providing a more reliable foundation for monocular 3D object detection.

Figure 3 illustrates the influence of multi-scale features on the performance of the attention mechanism. The figure presents the attention weight distributions across four different feature scales (from 1/32 to 1/4) and their corresponding 3D metrics. The results demonstrate that intermediate feature scales (1/16 and 1/8) receive higher attention weights (0.25 and 0.35, respectively) and simultaneously achieve better detection accuracy. This finding validates the design intention of incorporating multi-scale feature fusion that captures both global semantic information and fine-grained local details. In particular, the 1/8 scale feature contributes the most to accuracy improvement, providing an important reference for optimizing the allocation of attention weights in the model configuration.

Figure 4 compares the performance and performance of five different contextual coupling strategies. The adaptive coupling strategy showed the highest performance of 17.68%, but the longest withdrawal time was 1.2 milliseconds. The parallel synthesis strategy provides a relative balance between 17.80% accuracy and computation efficiency. This image clearly shows the relationship between different strategies that serves as the basis for choosing

Figure 4
Comparison of Different Context Fusion Strategies.



synthesis methods in practical applications. This analysis reflects the key importance of the contextual synthesis method to improve the performance of 3D detection and serves as an inspiration for later optimization of the model design.

4.3. Computational Complexity Analysis

To evaluate the practical applicability of the proposed method, its computational complexity is analyzed in terms of model parameters, floating point operations (FLOPs), and inference speed, and compared with representative monocular baselines.

All models are evaluated under the same experimental setting using an NVIDIA A6000 GPU and identical input resolution. The number of parameters and FLOPs are computed using the PyTorch profiling tool, while inference speed is measured in frames per second (FPS) with batch size set to one.

The results are summarized in Table X. Compared with MonoDETR, the proposed method introduces additional parameters due to the Global Self-Attention (GSA) and Local Context Fusion (LCF) modules. Specifically, the parameter count increases moderately, while the FLOPs exhibit a limited growth, mainly attributed to the attention-based global–local feature interaction.

Despite the increased computational cost, the inference speed of the proposed method remains within a practical range for autonomous driving applications. This indicates that the performance gains achieved by the proposed global–local context fusion strategy are obtained with an acceptable computational overhead, striking a reasonable balance between detection accuracy and efficiency.

As shown in Table 3, although the proposed method increases the computational overhead, the inference speed remains within a practical range, indicating a favorable trade-off between accuracy and efficiency.

Table 3
Computational complexity comparison with MonoDETR on the KITTI dataset.

Method	Params (M)	FLOPs (G)	FPS
MonoDETR	51.2	182.4	18.6
Proposed Method	56.8	205.7	16.9

5. Summary

This paper presents a monocular 3D object detection framework based on a multi-scale context fusion attention mechanism, aiming to address the challenges of depth ambiguity and insufficient global–local feature interaction in monocular settings.

The proposed method integrates a Global Self-Attention (GSA) module and a Local Context Fusion (LCF) module within a unified architecture. The GSA module captures long-range spatial dependencies and enhances global semantic consistency, while the LCF module focuses on preserving fine-grained local structures through similarity-aware neighborhood aggregation. By jointly modeling global scene context and local object details, the framework achieves more reliable depth reasoning and robust feature representation.

Extensive experiments conducted on the KITTI benchmark demonstrate that the proposed approach consistently outperforms representative monocular 3D detection baselines, particularly in challenging scenarios involving distant and partially occluded objects. Moreover, the computational complexity analysis shows that the performance gains are achieved with a moderate and practical computational overhead.

Overall, the proposed global–local context fusion strategy provides an effective and scalable solution for monocular 3D object detection, offering a favorable balance between accuracy, robustness, and efficiency in autonomous driving applications.

References

1. Alaba, Y. S., Gurbuz, C. A., Ball, E. J. Emerging Trends in Autonomous Vehicle Perception: Multimodal Fusion for 3D Object Detection. World Electric Vehicle Journal, 2024, 15(1), 20. https://doi.org/10.3390/wevj15010020
2. Alexandre, E., Moliard, A., Grzeskowiak, F., Vair, C. Improving the Efficiency of 3D Monocular Object Detection and Tracking for Road and Railway Smart Mobility. Sensors, 2023, 23(6), 3197. https://doi.org/10.3390/s23063197
3. Amarasiri, N. M. B. A., Silva, D. A. L. C., Fernando, L. Impact of LiDAR Point Cloud Compression On 3D Object Detection Evaluated on the KITTI dataset. EURASIP Journal on Image and Video Processing, 2024, 2024(1), 11. https://doi.org/10.1186/s13640-024-00633-4
4. Boyko, N. Development of an Efficient Method for Object Detection and Localization in 3D Space Using RGBD Cameras for Autonomous Systems. International Journal of Cognitive Computing in Engineering, 2025, 6, 537-551. https://doi.org/10.1016/j.ijcce.2024.11.002
5. Contreras, M., Jain, A., Bhatt, P. N., Liu, Y. H. A Survey on 3D Object Detection in Real Time for Autonomous Driving. Frontiers in Robotics and AI, 2024, 11, 1212070. https://doi.org/10.3389/frobt.2024.1212070
6. Dhakal, S., Qu, D., Carrillo, D., Chen, Y., Liu, Y. VirtualPainting: Addressing Sparsity with Virtual Points and Distance-Aware Data Augmentation for 3D Object Detection. Sensors, 2025, 25(11), 3367. https://doi.org/10.3390/s25113367
7. Ding M, Huo Y, Yi H, Wang Z, Shi J, Lu Z, Luo P. Learning Depth-Guided Convolutions for Monocular 3D Object Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020. https://doi.org/10.1109/CVPR42600.2020.01169
8. Duy, L., Linh, N. Simple Linear Iterative Clustering Based Low-Cost Pseudo-Lidar for 3D Object Detection in Autonomous Driving. Multimedia Tools and Applications, 2023, 82(16), 25253-25269. https://doi.org/10.1007/s11042-023-14439-5
9. Erabati, K. G., Araujo, H. DeLiVoTr: Deep and light-Weight Voxel Transformer for 3D Object Detection. Intelligent Systems with Applications, 2024, 22, 200361. https://doi.org/10.1016/j.iswa.2024.200361
10. Erabati, K. G., Araujo, H. SRFDet3D: Sparse Region Fusion based 3D Object Detection. Neurocomputing, 2024, 593, 127814. https://doi.org/10.1016/j.neucom.2024.127814
11. Essien, E., Frimpong, S. Enhancing Autonomous Truck Navigation in Underground Mines: A Review of 3D Object Detection Systems, Challenges, and Future Trends. Drones, 2025, 9(6), 433. https://doi.org/10.3390/drones9060433
12. Fawole, A. O., Rawat, B. D. Recent Advances in 3D Object Detection for Self-Driving Vehicles: A Survey. AI, 2024, 5(3), 1255-1285. https://doi.org/10.3390/ai5030061
13. Ghazal, R., Challita, M. A., Charpillet, F., Pu, P. A Deep Model of Visual Attention for Saliency Detection on 3D Objects. Neural Processing Letters, 2023, 55(7), 8847-8867. https://doi.org/10.1007/s11063-023-11180-w
14. Housam, P. S., Manuel, J. V. R., Louahdi, K., Asier, A., Oihana, O., Marcos, M. 3D Object Detection for Self-Driving Cars Using Video and LiDAR: An Ablation Study. Sensors, 2023, 23(6), 3223. https://doi.org/10.3390/s23063223
15. Huang, K. C., Wu, T. H., Su, H. T., Hsu, W. H. MonoDTR: Monocular 3D Object Detection with Depth-Aware Transformer. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022. https://doi.org/10.1109/CVPR52688.2022.00398
16. Yan, L., Ma, X., Liu, Y., Zhang, T., Liu, Y., Chu, Q. MonoCD: Monocular 3D Object Detection with Complementary Depths. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024. https://doi.org/10.1109/CVPR52733.2024.00976
17. Yildiz, S. A., Meng, H., Swash, R. M. Real-Time Object Detection and Distance Measurement Enhanced with Semantic 3D Depth Sensing Using Camera-LiDAR Fusion. Applied Sciences, 2025, 15(10), 5543. https://doi.org/10.3390/app15105543
18. Jaber, N., Wehbe, B., Kirchner, F. 3D-DUO: 3D Detection of Underwater Objects in Low-Resolution Multibeam Echosounder Maps. Ocean Engineering, 2025, 331, 121254. https://doi.org/10.1016/j.oceaneng.2025.121254
19. Joshi, R., Usmani, K., Krishnan, G., Watson, P., Javidi, B. Underwater Object Detection and Temporal Signal Detection in Turbid Water Using 3D-Integral Imaging and Deep Learning. Optics Express, 2024, 32(2), 1789-1801. https://doi.org/10.1364/OE.510681
20. Kozov, V., Minev, E., Andreeva, M., Milchev, M., Georgiev, S. Comparative Analysis of Different Display Technologies for Defect Detection in 3D Objects. Technologies, 2025, 13(3), 118. https://doi.org/10.3390/technologies13030118
21. Liu, X., Xue, N., Wu, T. Learning Auxiliary Monocular Contexts Helps Monocular 3D Object Detection. Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). 2022. https://doi.org/10.1609/aaai.v36i2.20074

22. 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. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2021. https://doi.org/10.1109/ICCV48922.2021.00310

23. Ma, X., Zhang, Y., Xu, D., Zhou, D., Yi, S., Li, H., Ouyang, W. Delving into Localization Errors for Monocular 3D Object Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021.

24. Naich, Y. A., Carrión, R. J. LiDAR-Based Intensity-Aware Outdoor 3D Object Detection. Sensors, 2024, 24(9), 2942. https://doi.org/10.3390/s24092942

25. Noh, J., Lee, J., Park, H., Lee, D., Kweon, I. S. 3DPillars: Pillar-based Two-Stage 3D Object Detection. Expert Systems With Applications, 2025, 289, 128349. https://doi.org/10.1016/j.eswa.2025.128349

26. Nurfauzi, R., Baba, A., Nakada, A. T., Murata, E., Mori, K. Automated Traumatic Bleeding Detection in Whole-Body CT Using 3D Object Detection Model. Applied Sciences, 2025, 15(15), 8123. https://doi.org/10.3390/app15158123

27. Park, G., Koh, J., Kim, J., Lee, S. LiDAR-Based 3D Temporal Object Detection via Motion-Aware LiDAR Feature Fusion. Sensors, 2024, 24(14), 4667. https://doi.org/10.3390/s24144667

28. Qin, Z., Li, X. MonoGround: Detecting Monocular 3D Objects from the Ground. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022. https://doi.org/10.1109/CVPR52688.2022.00377

29. Ramana, R., Vasudevan, V., Murugan, S. B. Spectral Pyramid Pooling and Fused Keypoint Generation in ResNet-50 for Robust 3D Object Detection. IETE Journal of Research, 2025, 71(4), 1220-1232. https://doi.org/10.1080/03772063.2025.2453897

30. Reading, C., Harakeh, A., Chae, J., Waslander, S. L. Categorical Depth Distribution Network for Monocular 3D Object Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021. https://doi.org/10.1109/CVPR46437.2021.00845

31. Reuse, M., Amende, K., Simon, M., Kummert, A. Exploring 3D Object Detection for Autonomous Factory Driving: Advanced Research on Handling Limited Annotations with Ground Truth Sampling Augmentation. Computer Sciences & Mathematics Forum, 2024, 9(1), 5. https://doi.org/10.3390/cmsf2024009005

32. Shi, P., Jiang, T., Yang, A., Xie, P., Zheng, Z., Yang, Y. Cen-Radfusion: Fusing Image Center Detection and Millimeter Wave Radar for 3D Object Detection. Signal, Image and Video Processing, 2024, 18(8-9), 5811-5821. https://doi.org/10.1007/s11760-024-03273-3

33. Soans, R., Fukumizu, Y. Custom Anchorless Object Detection Model for 3D Synthetic Traffic Sign Board Dataset with Depth Estimation and Text Character Extraction. Applied Sciences, 2024, 14(14), 6352. https://doi.org/10.3390/app14146352

34. Tiwari, K. A., Sharma, K. G. FS-3DSSN: An Efficient Few-Shot Learning for Single-Stage 3D Object Detection on Point Clouds. The Visual Computer, 2024, 40(11), 8125-8139. https://doi.org/10.1007/s00371-023-03228-8

35. Usmani, K., O'Connor, T., Watson, P., Javidi, B. 3D Object Detection Through Fog and Occlusion: Passive Integral Imaging vs Active (Lidar) Sensing. Optics Express, 2023, 31(1), 479-491. https://doi.org/10.1364/OE.478125

36. Vosselman, G., Xiao, Y. Change Detection of Urban Objects Using 3D Point Clouds: A Review. ISPRS Journal of Photogrammetry and Remote Sensing, 2023, 197, 228-255. https://doi.org/10.1016/j.isprsjprs.2023.01.010

37. Zakeri, A., Koirala, B., Kang, J., Kim, C. S. SMS3D: 3D Synthetic Mushroom Scenes Dataset for 3D Object Detection and Pose Estimation. Computers, 2025, 14(4), 128. https://doi.org/10.3390/computers14040128

38. Zamanakos, G., Tsochatzidis, L., Amanatiadis, A., Gasteratos, A. Feature Aware Re-weighting (FAR) in Bird's Eye View for LiDAR-based 3D object detection in autonomous driving applications. Robotics and Autonomous Systems, 2024, 175, 104664. https://doi.org/10.1016/j.robot.2024.104664

39. Zhang ,Y., Lu, J., Zhou, J. Objects are Different: Flexible Monocular 3D Object Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021. https://doi.org/10.1109/CVPR46437.2021.00330

40. Zhang, R., Qiu, H., Wang, T., Guo, Z., Tang, Y., Xu, X., Cui, Z., Qiao, Y., Gao, P., Li, H. MonoDETR: Depth-guided Transformer for Monocular 3D Object Detection. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2023. https://doi.org/10.1109/ICCV51070.2023.00840