Transactions of KSAE, Vol. 34, No. 9, pp.1009-1019 (September, 2026)

pISSN 1225-6382 / eISSN 2234-0149
DOI http://dx.doi.org/10.7467/KSAE.2026.34.9.1009

Segmentation Residual–Gated Multi-Scale Region Enhancement in Monocular 3D Object Detection

Tao Peng · Jinsu An · Byeong Woo Kim*

Department of Electrical, Electronic and Computer Engineering, University of Ulsan, Ulsan 44610, Korea

(Received 29 January 2026 / Revised 17 April 2026 / Accepted 21 April 2026)

Abstract : Monocular 3D object detection faces particular challenges with vulnerable road users (VRUs) due to their small scale, frequent occlusions, and unreliable long-range depth cues. The depth-guided transformer was enhanced to better model VRU geometry. First, projection-based geometric depth is treated as an independent prior while a geometry-error residual is learned, which stabilizes optimization for distant VRUs. Second, we decouple the 2D and 3D heads so the 2D head handles detection and query initialization, while the 3D head concentrates on depth, center, size, and orientation, thereby reducing gradient interference. Third, we introduce query-adaptive gated depth cross-attention to selectively utilize depth features and suppress ambiguous cues in difficult cases. We reinforce region conditioning through an upgraded region segmentation head module that incorporates learnable multi-scale fusion, channel–spatial attention, and residual gating. Experiments on the KITTI test set demonstrate consistent gains for cars, pedestrians, and cyclists across all difficulty splits.

Key words : Monocular 3D object detection, Vulnerable road users, Depth-guided transformer, Query-adaptive gated cross-attention, Residual geometry correction

1. Introduction

Monocular 3D object detection is a critical enabling technology for vision-based perception in autonomous driving and advanced driver assistance systems. It offers a cost-effective alternative to expensive 3D sensors.[1–4] Unlike LiDAR-based approaches, monocular methods[5–8] must infer 3D structure from a single RGB image with camera intrinsics, introducing ambiguity in depth estimation, scale recovery, and occlusion reasoning. Recently, transformer-based detectors[9–11] have advanced the field by jointly modeling appearance and geometry through end-to-end query-based frameworks. Representative methods such as MonoDETR[12] and MonoDGP[13] demonstrate that depth-guided transformers can effectively exploit geometric priors and learned depth features for monocular 3D detection.

Despite progress, performance is fragile for vulnerable road users(VRUs) like pedestrians and cyclists. Cars dominate training, allowing geometric priors and depth supervision to be learned primarily from car-rich regimes. In contrast, VRUs are sparsely observed beyond medium range, exposing the model to limited depth variation during training. Under this bias, directly enforcing depth-related losses can overfit to car-dominant statistics and impose overly rigid constraints that don't generalize well to distant or partially occluded VRUs. Moreover, coupling 2D object detection and 3D regression within shared prediction heads may restrict task specialization. Consequently, improvements aimed at challenging cases (VRUs, far-range objects) are often limited.

Building upon MonoDGP, we introduce the following key contributions:

- We treat projection-based geometric depth as a detached prior and supervise a learnable geometry-error residual, stabilizing optimization and improving robustness for distant objects and VRUs.
- We decouple the 2D and 3D heads so that 2D focuses on detection and query initialization, while 3D specializes in depth, center, size, and orientation regression. This

*Corresponding author, E-mail: bywokim@ulsan.ac.kr

design reduces optimization interference between image-plane perception and 3D geometric reasoning, and provides a more suitable feature basis for subsequent depth-aware fusion.

• We introduce a gated depth cross-attention module in the 3D decoder, enabling per-query modulation of depth feature contribution—down-weighting unreliable cues in hard/long-range cases while exploiting strong depth evidence when available.

• We upgrade the region enhancement module with learnable multi-scale fusion, channel–spatial attention, and residual gating to preserve informative features and provide more stable depth-conditioned representations, benefiting VRU detection across difficulty levels.

2. Related Works

Transformer based 3D Object Detection: Transformer based detectors have revolutionized object detection by enabling end-to-end set prediction without the need for dense proposal generation or heuristic post-processing. The seminal work introduced DETR,[15] which predicts a fixed-size set of objects using a transformer encoder–decoder with learned object queries. By replacing anchor design and non-maximum suppression with bipartite matching, DETR provides a clean optimization objective and produces object-centric representations that are iteratively refined through attention. This query-centric paradigm is particularly attractive for monocular 3D object detection, as it naturally supports object-level reasoning and facilitates the fusion of heterogeneous cues via cross-attention.

Inspired by DETR, early transformer based monocular 3D detectors adopt query-based reasoning to regress 3D parameters in an end-to-end manner. MonoDETR further strengthens this by explicitly integrating depth guidance into the transformer pipeline. It uses a depth prediction module to provide depth-aware embeddings fused with visual features, aligning object queries with 3D structure and reducing ambiguities from similar 2D evidence. This improves robustness under viewpoint variations and partial occlusions by providing additional geometric constraints.

MonoDGP (Fig. 1) extends depth-guided transformer detection with two notable ideas: decoupled query reasoning and geometry-aware depth modeling. Instead of a single decoder jointly handling image-plane detection and 3D inference, MonoDGP first forms 2D hypotheses then refines them for 3D localization using depth cues. This staged design promotes specialization—appearance-based localization in the 2D branch and metric reasoning under depth uncertainty in the 3D branch. Moreover, MonoDGP explicitly models depth using a pinhole-camera formulation and object statistics, introducing a learnable residual term with uncertainty-aware supervision to correct systematic geometric bias.

However, several issues continue to hinder robust performance on long-range objects and VRUs. First, geometric depth signals are often tightly coupled with 2D box geometry and 3D size priors, such that depth-related supervision may inadvertently reshape dimensions or influence object scale during optimization, which can destabilize training and amplify car-dominant biases in imbalanced datasets. Second, even when 2D and 3D reasoning are decoupled, partially shared prediction components can still entangle gradients, limiting specialization for small and distant VRUs. Third, depth features are typically injected uniformly, despite varying reliability; noisy depth cues for far-range VRUs may mislead the decoder and reduce gains on Hard settings.

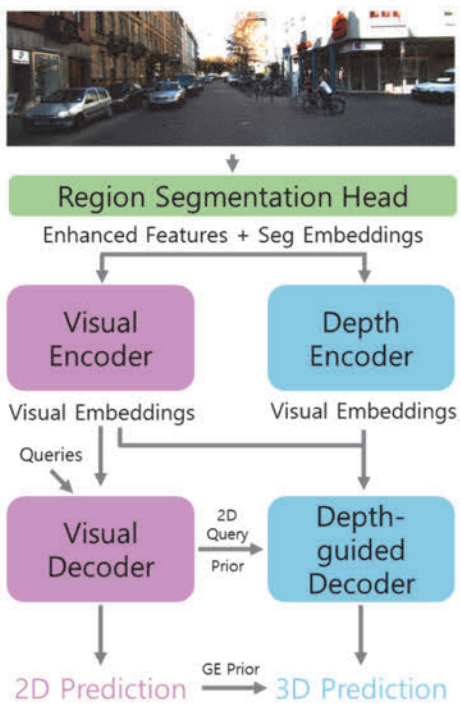

Fig. 1 Overview of the MonoDGP architecture

3. Methodology

Building upon the foundational MonoDGP pipeline, we introduce a stable, selective, and geometry-aware depth guidance mechanism. This enhancement directly tackles the critical challenge of unreliable depth cues in monocular 3D object detection.

3.1 Overall Architecture

Given an input image I of size (H, W), a CNN backbone extracts multi-scale visual features F_l. In parallel, a depth predictor produces depth-related representations to guide 3D reasoning. Unlike the original MonoDGP, we strengthen this guidance with an enhanced region segmentation head that performs learnable multi-scale fusion, channel–spatial attention, and residual-gated feature enhancement. This design highlights object-relevant regions while avoiding over-suppression when the region confidence is uncertain, thereby providing more stable depth-conditioned cues. The refined depth representations are then encoded by a depth encoder into multi-scale depth embeddings.

3.2 Enhanced Region Segmentation Module

MonoDGP employs a Region Segmentation Module (RSM) to estimate target-region probabilities from multi-scale features and enhance depth-related representations by suppressing background responses. Although effective, the original RSM relies on a fixed top-down fusion strategy and direct multiplicative gating, which may allow noisy high-level responses to dominate feature aggregation, lack sufficient spatial selectivity in cluttered street scenes, and over-suppress informative features when the predicted region probability is uncertain. These limitations are particularly detrimental to small and long-range VRUs, whose visual evidence is weak and easily overwhelmed by background structures.

To address these issues, we propose an Enhanced Region Segmentation Module (E-RSM), which improves the original RSM through three targeted upgrades: learnable weighted fusion, channel-spatial attention, and residual gated enhancement. The goal is to obtain more stable region-aware features while preserving weak but informative cues for downstream monocular 3D detection.

3.2.1 Learnable Weighted Multi-scale Fusion

In the original RSM, the current-level feature and the upsampled higher-level feature are fused with fixed equal weighting. Such a design is simple but may cause overconfident semantic responses from high-level features

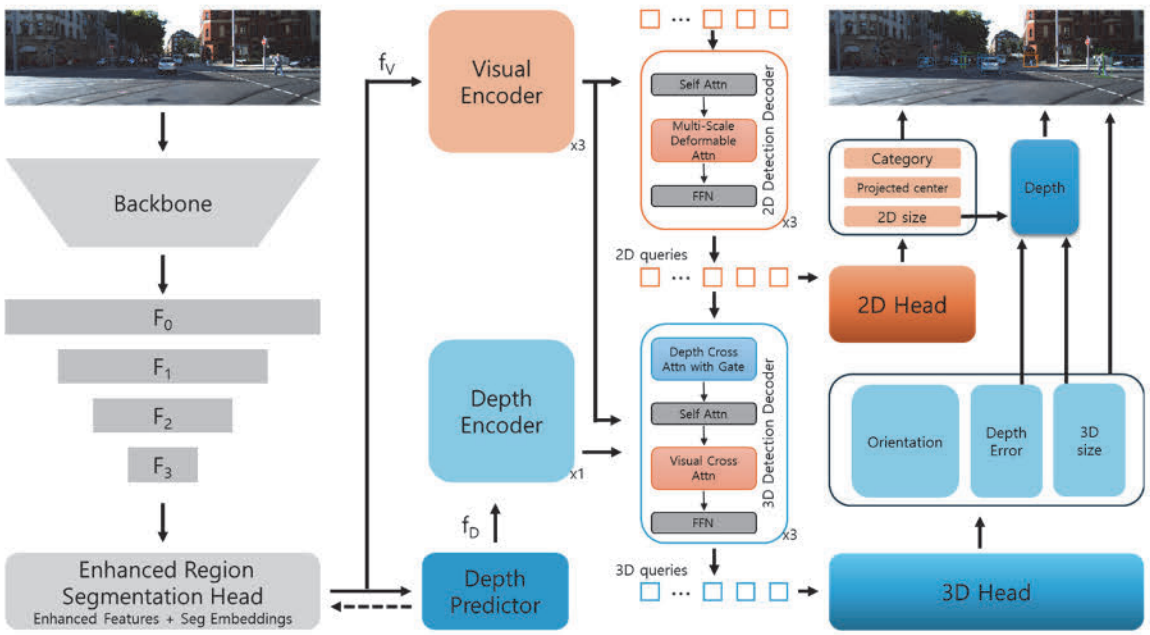

Fig. 2 Overview of our optimized MonoDGP framework. The model comprises multi-scale feature extraction, a decoupled 2D/3D transformer pipeline with depth guidance, and task-specific prediction heads

to dominate lower-level representations, especially when the target object is small, distant, or partially occluded.

To alleviate this issue, E-RSM introduces learnable fusion weights for each top-down connection. This allows the network to adaptively balance coarse semantic context and fine spatial details according to the input feature characteristics. As a result, the fusion process becomes less sensitive to noisy high-level activations and more capable of preserving weak object evidence, which is particularly important for VRUs under challenging street-scene conditions.

3.2.2 Channel + Spatial Attention Refinement

After multi-scale fusion, E-RSM applies a two-stage attention refinement consisting of channel attention followed by spatial attention. The channel attention stage reweights feature channels according to their importance, while the spatial attention stage emphasizes target-aligned regions and suppresses background-dominant responses.

This refinement is particularly useful in urban scenes, where road markings, fences, poles, and building edges often produce strong but misleading activations. By explicitly modeling both what to emphasize and where to emphasize it, E-RSM improves region selectivity and helps the detector focus on object-supportive areas rather than structured background clutter. This is especially beneficial for small-scale pedestrians and cyclists.

3.2.3 Residual Gated Enhancement

The original RSM enhances features through direct multiplicative modulation (feature × probability). Although intuitive, this strategy can unintentionally suppress useful information when the predicted region probability is uncertain, which is common for distant or weakly visible objects.

To improve robustness, E-RSM replaces direct multiplication with a residual gated enhancement scheme. Instead of using the probability map to overwrite the original feature response, the original feature is preserved as a stable base representation, while the region probability provides an additional enhancement term. In this way, confident target regions are strengthened, but informative features are not excessively attenuated when the region confidence is low.

This residual-style design makes the feature enhancement process less brittle and provides the downstream transformer with more stable region-aware representations, thereby reducing missed detections caused by overly aggressive suppression.

3.3 Detached Geometric Depth Prior

We introduce the Geometric Depth Prior module to enable geometrically consistent depth estimation. Leveraging the pinhole camera model, this module computes a geometric depth proxy, Z_{geo}, exploiting the correlation between an object's physical dimensions and its projected size. The formulation is defined as:

$$Z_{geo} = \frac{f \cdot H_{3D}}{h_{bbox}} \tag{1}$$

where f is the focal length, H_{3D} is the predicted 3D physical height, and h_{bbox} is the height of the 2D bounding box (clamped for numerical stability).

Crucially, we apply a stop-gradient operation on inputs H_{3D} and h_{bbox} before computation. This decouples depth estimation from the dimension regression task, ensuring that Z_{geo} serves as a stable reference signal without interference from the height branch's learning. By preventing backpropagation of gradients through this geometric path, we improve overall training stability.

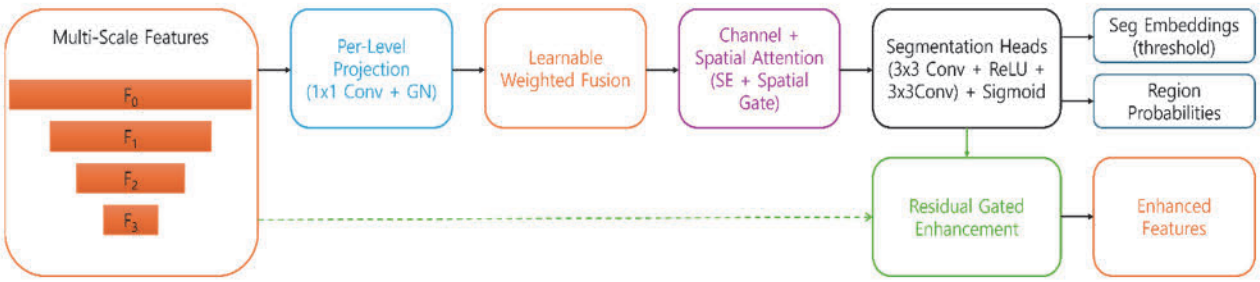

Fig. 3 Simplified overview of the enhanced Region Segmentation Module (E-RSM)

3.4 Decoupled 2D 3D Prediction Heads

To reduce feature interference between 2D perceptual tasks and 3D geometric reasoning, we adopt a decoupled task head architecture. Unlike shared-head designs, separate projection layers are employed for different query stages. The 2D head processes Q_{2D} from the 2D decoder to produce classification scores and normalized 2D boxes, while the 3D head processes Q_{3D} from the 3D decoder to regress 3D centers, dimensions, depths, and orientations. This structural separation encourages task-specific representation learning for image-plane detection and 3D geometry estimation, alleviates optimization conflict, and provides a cleaner feature basis for the subsequent depth-guided fusion module.

3.5 Adaptive Depth-Gated Fusion

To improve the incorporation of geometric information within the 3D decoder, we introduce an Adaptive Depth-Gated Mechanism inside the Depth Aware Decoder Layer. In conventional Transformer designs, depth features obtained via cross-attention are usually fused directly into object queries through residual addition. However, not all depth tokens provide accurate or pertinent cues for every object query. To address this, we implement a learnable gating module to regulate feature injections. Let Q denote the current object query and F_{depth} denote the depth features extracted via cross-attention. We compute a modulation gate $G \in (0, 1)$ using a lightweight Multi-Layer Perceptron applied to Q:

$$G = \sigma(MLP(Q)) \tag{2}$$

where σ is the Sigmoid activation function, ensuring the gate values are between 0 and 1. This gate has the same dimensionality as the query, enabling fine-grained, channel-wise control. The depth information is then fused back into the query stream via a gated residual connection:

$$Q' = LayerNorm(Q + Dropout(G \odot F_{depth})) \tag{3}$$

By using element-wise multiplication, the network can adaptively amplify high-confidence geometric cues while suppressing noise or irrelevant depth information for specific queries, stabilizing the 3D reasoning process.

4. Experiments

4.1 Dataset and Evaluation Metrics

We evaluate our method on the widely used KITTI[14] 3D object detection benchmark. The dataset contains 7,481 training images and 7,518 test images, covering three categories (Car, Pedestrian, and Cyclist) and three difficulty levels (Easy, Moderate, and Hard). Following common practice,[16] we split the 7,481 training images into 3,712 images for training and 3,769 images for validation to conduct ablation studies and model selection.

We report detection performance using the official KITTI protocol[17] with average precision computed at 40 recall positions ($AP|_{R40}$). For Cars, AP_{3D} is evaluated at IoU = 0.7, while Pedestrian and Cyclist are evaluated at IoU = 0.5. Unless stated otherwise, all models are trained using the same settings as the MonoDGP baseline to ensure fair comparison.

4.1.1 Main Result

Table 1 summarizes the KITTI test-set results for Car, Pedestrian, and Cyclist. Compared with MonoDGP, our method achieves consistent improvements across all three categories, with particularly strong gains on VRUs. For the Car category, our approach improves MonoDGP from 26.35/18.72/15.97 to 26.74/18.96/16.26 on Easy/Mod./Hard, corresponding to +0.41/+0.24/+0.29 AP_{3D}.

More importantly, our method yields substantial improvements for VRUs, which are more sensitive to depth noise and long-range ambiguity. For Pedestrian, we improve MonoDGP from 15.04/9.89/8.38 to 16.96/10.90/9.36, achieving +1.92/+1.01/+0.98 gains on Easy/Moderate/Hard. For Cyclist, our method improves from 5.28/2.82/2.65 to 8.35/4.80/3.66, yielding even larger gains of +3.07/ +1.98/+1.01. These results indicate that our geometry-stabilized depth supervision and query-adaptive depth usage substantially reduce depth-induced errors for small and distant VRUs.

Table 2 further confirms these trends on the KITTI validation split. Relative to MonoDGP, our method improves Car to 22.55 / 19.55 on Moderate / Hard (from 22.34 / 19.02), and yields clear gains for VRUs, raising Pedestrian from 10.06 to 11.56 and Cyclist from 6.61 to 7.66 on the Moderate setting. While Car Easy slightly drops (30.76 → 30.13), the overall improvements on

Table 1 The $AP|_{R40}$ scores on the KITTI test set for 3D object detection in the Car, Pedestrian, and Cyclist categories. **Bold** text indicates the best results for each category, while underlined text represents the second-best results. *Red* highlights indicate improvements versus MonoDGP

Methods	Extra data	Car, AP_{3D}@IOU=0.7			Pedestrian, AP_{3D}@IOU=0.5			Cyclist, AP_{3D}@IOU=0.5		
		Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
MonoRUn[18]	LiDAR	19.65	12.30	1058	11.18	6.53	5.73	0.69	0.55	0.42
CaDDN[19]		19.17	13.41	11.46	12.87	8.14	6.76	7.00	3.41	3.30
MonoDTR[9]		21.99	15.39	12.73	15.33	10.18	8.61	5.05	3.27	3.19
D4LCN[20]	Depth	16.65	11.72	9.51	4.55	3.42	2.83	2.45	1.67	1.36
DDMP[21]		19.71	12.78	9.80	4.93	3.55	3.01	4.18	2.50	2.32
MonoPair[22]	None	13.04	9.99	8.65	10.02	6.68	5.53	3.79	2.12	1.83
M3DSSD[24]		17.51	11.46	8.98	5.16	3.87	3.08	2.10	1.51	1.58
MonoDLE[25]		17.23	12.26	10.29	9.64	6.55	5.44	4.59	2.66	2.45
MonoFlex[26]		19.94	13.89	12.07	9.43	6.31	5.26	4.17	2.35	2.04
DEVIANT[23]		21.88	14.46	11.89	13.43	8.65	7.69	5.05	3.13	2.59
MonoDETR		25.00	16.47	13.58	12.54	7.89	6.65	7.33	4.18	2.92
MonoDGP		26.35	18.72	15.97	15.04	9.89	8.38	5.28	2.82	2.65
Ours	None	**26.74**	**18.96**	**16.26**	**16.96**	**10.90**	**9.36**	**8.35**	**4.80**	**3.66**
Improvement		*+0.41*	*+0.24*	*+0.29*	*+1.92*	*+1.01*	*+0.98*	*+3.07*	*+1.98*	*+1.01*

Table 2 $AP|_{R40}$ scores on the KITTI validation set for 3D object detection in the Car, Pedestrian and Cyclist categories

Methods	Extra data	Car, AP_{3D}@IOU=0.7			Pedestrian, AP_{3D}@IOU=0.5			Cyclist, AP_{3D}@IOU=0.5		
		Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
D4LCN[20]	Depth	26.97	21.71	18.22	12.95	11.23	11.05	5.85	4.41	4.14
DDMP[21]		28.12	20.39	16.34	14.42	**12.11**	**12.05**	8.01	6.47	6.27
DID-M3D[27]	None	22.98	16.12	14.03	7.27	5.87	4.89	5.54	2.59	2.49
GUPNet[28]		22.76	16.46	13.72	9.37	6.84	5.73	4.41	2.17	2.03
DEVIANT		24.63	16.54	14.52	9.85	7.18	5.42	4.05	2.20	2.14
MonoDGP		**30.76**	22.34	19.02	13.77	10.06	7.96	12.21	6.61	5.95
Ours	None	30.13	**22.55**	**19.55**	**15.15**	11.56	9.10	**15.53**	**7.66**	**7.22**
Improvement		-0.63	+0.21	+0.53	+1.38	+1.50	+1.14	+3.32	+1.05	+1.27

Moderate/Hard and on VRU categories suggest that the proposed depth guidance is particularly effective in challenging scenarios where depth cues are less reliable and small-object geometry is harder to recover.

4.1.2 Multi-seed Robustness Analysis

To examine whether the previously observed -0.63 drop on Car Easy in a single run reflects a structural side effect of the proposed modules or merely random initialization variance, we further conducted a multi-seed robustness analysis using five random seeds: 42, 444, 1024, 2024, and 3407. Among them, 444 corresponds to the fixed random seed adopted in the original MonoDGP implementation.

The results are summarized in Table 3. On car AP_{3D}, the baseline MonoDGP achieves 29.10 ± 0.34 / 21.17 ± 0.46 / 18.19 ± 0.56 on Easy / Moderate / Hard, respectively, whereas the proposed method achieves 29.24 ± 1.07 / 21.57 ± 0.58 / 18.74 ± 0.54. Therefore, the previously observed single-run drop on car easy is not consistently reproduced under multiple random seeds. After averaging across seeds, the proposed method shows a slight improvement on Easy (+0.14) and clearer gains on Moderate (+0.39) and Hard (+0.55).

A paired seed-wise comparison further supports this observation. Relative to MonoDGP, the proposed method improves 3/5 runs on Easy and 4/5 runs on both Moderate

Table 3 Multi-seed robustness analysis of MonoDGP and the proposed method on the KITTI validation set

Methods	Seed	Car. AP_{3D} @IOU=0.7		
		Easy	Mod.	Hard
Baseline	42	29.12	20.52	17.30
	444	29.62	21.33	18.33
	1024	28.93	21.33	18.31
	2024	28.70	20.96	18.17
	3407	29.12	21.73	18.83
Ours	42	27.62	21.06	18.29
	444	30.13	22.55	19.55
	1024	29.88	21.40	18.38
	2024	29.69	21.24	19.02
	3407	29.89	21.59	18.45

Table 4 Ablation on different components

Methods	Mod. AP_{3D}		
	Car, @IOU=0.7	Pedestrian @IOU=0.5	Cyclist @IOU=0.5
Baseline	21.24	9.57	5.94
Baseline+DDP	21.52	9.40	7.62
Baseline+DDP +DPH	20.78	10.32	6.60
Baseline+DDP +DPH+ADG	21.65	10.82	8.94
Baseline+DDP +ADG+E-RSM	21.21	10.61	8.43
Baseline+DDP +DPH+ADG+E-RSM	22.55	11.56	7.66

and Hard. The remaining decreases are limited to a few cases and are small in magnitude, indicating that the earlier Easy drop is better interpreted as seed-dependent variance rather than a systematic structural drawback of the proposed design.

Overall, these results demonstrate that the proposed mo dules do not introduce a consistent negative effect on car e asy, while providing more stable improvements on the mo re challenging Moderate and Hard subsets.

4.2 Ablation Study

4.2.1 Ablation on Different Components

We conduct an ablation study on the KITTI validation set using Moderate AP3D for Car, Pedestrian, and Cyclist, as summarized in Table 4. Here, DDP denotes the detached geometric depth prior, DPH denotes the decoupled 2D/3D prediction heads, ADG denotes adaptive depth-gated fusion, and E-RSM denotes the enhanced region segmentation module.

The results reveal several important trends. First, DDP improves Car and Cyclist over the baseline, indicating that detached geometric supervision provides a useful and stable geometric cue. Second, adding DPH on top of DDP does not yield uniform gains across all categories: it improves Pedestrian, but decreases Car and Cyclist in this intermediate setting. This suggests that the benefit of DPH is not a standalone monotonic gain, but rather a structural decoupling effect whose value becomes more evident when combined with later depth-aware modules. Third, after introducing ADG, performance improves markedly on all three categories, showing that query-wise adaptive depth gating is the key component that effectively converts the decoupled 3D branch into stronger geometric reasoning.

To further address whether DPH is necessary, we additionally report the variant without DPH, i.e., DDP+ ADG+E-RSM. Compared with this setting, the full model DDP+DPH+ADG+E-RSM improves Car from 21.21 to 22.55 and Pedestrian from 10.61 to 11.56, while Cyclist decreases from 8.43 to 7.66. Therefore, DPH introduces a category-dependent trade-off rather than a uniform gain. Nevertheless, it contributes positively to the final overall design by improving the two more critical categories in our setting, namely Car and Pedestrian, and by working complementarily with ADG and E-RSM in the full model.

4.2.2 Ablation on Inference Speed

To clarify whether the gain of the proposed E-RSM simply comes from inserting a generic attention module, we further compare it with two CBAM-based alternatives:

(1) + CBAM, which directly replaces the enhancement stage with a standard CBAM module, and

(2) Learnable weighted fusion + CBAM + residual gating, which combines several related components in a straightforward manner.

The comparison results are shown in Table 5. In contrast to these baselines, the proposed E-RSM is not a simple stacking of generic modules. Instead, it is designed as a task-oriented unified region enhancement module, in which adaptive multi-scale fusion, channel-spatial refinement, and residual gated modulation are jointly organized within a single residual enhancement pathway.

Table 5 Quantitative comparison between CBAM-based variants and the proposed E-RSM on the KITTI validation set

Methods	Mod. AP_{3D}		
	Car, @IOU=0.7	Pedestrian @IOU=0.5	Cyclist @IOU=0.5
Baseline	21.24	9.57	5.94
+ CBAM	21.09	8.51	4.7
+ Learnable weighted fusion + CBAM + residual gating	21.41	8.8	6.5
Ours	22.55	11.56	7.66

Quantitatively, the representative result of the proposed method reported in Table 5 outperforms Learnable weighted fusion + CBAM + residual gating, the proposed E-RSM achieves better performance on nearly all metrics, demonstrating that its effectiveness cannot be attributed to a simple combination of generic components.

The improvements are especially clear on Pedestrian and Cyclist, suggesting that E-RSM is more effective at preserving weak foreground evidence and suppressing structured background clutter than generic attention insertion. These results indicate that the gain of E-RSM does not arise merely from adding CBAM-like attention, but from its integrated task-specific design for region-aware feature enhancement.

4.2.3 Ablation on Inference Speed

Table 6 reports the computational cost and inference latency measured on a single NVIDIA A100 GPU with batch size 1 under the same codebase and software environment. All results are obtained with an input resolution of 384×1280 using 50 warm-up iterations followed by 5 repeated runs of 100 timed iterations, and latency is measured with synchronized CUDA events. Although our model increases the parameter count and FLOPs from 38.90 M / 68.96 G to 83.01 M / 140.18 G, it achieves lower end-to-end latency, reducing the runtime from 42.96 ± 1.88 ms to 35.33 ± 5.57 ms. This result indicates that practical inference speed is not determined solely by FLOPs, but also by operator efficiency and execution patterns on GPU. Profiling further shows that both models are mainly dominated by convolutional operators, while our model shifts a larger portion of computation to highly optimized cuDNN convolution kernels; meanwhile, the relative cost of attention-related operators such as matrix multiplication, softmax, and deformable attention does not increase proportionally. At the same time, tensor-layout conversion and frequent element-wise operations remain non-negligible bottlenecks in both models, suggesting room for further efficiency optimization.

Table 6 Ablation study of computational cost. We test the Runtime (ms) on a single A100 GPU with a batch size of 1

Methods	Params (M)↓	FLOPs (G)↓	Runtime (ms)↓
MonoDGP	38.90	68.99	42.96±1.88
Ours	83.01	140.18	35.33±5.57

4.3 Visualization

Fig. 4 compares qualitative results on the KITTI validation set between MonoDGP and ours. Our predictions exhibit cleaner 3D box alignment and more consistent depth/scale estimation, particularly for VRUs under long-range and occluded conditions. In contrast to MonoDGP, our method generates tighter, more coherent cyclist/pedestrian boxes with fewer spurious detections, indicating that query-adaptive depth gating effectively suppresses unreliable depth cues in difficult regions. Ours method better preserves object extent and orientation, yielding boxes that more closely match the true geometry when cyclists are partially occluded or adjacent to vehicles.

Finally, in the long-range cases, our method demonstrates improved depth consistency and reduced false positives around background structures, leading to more reliable 3D localization in both the camera view. These qualitative observations align with our quantitative gains, confirming that stabilizing and selectively trusting depth guidance is crucial for robust VRU-oriented monocular 3D detection.

5. Conclusions

This paper presents a set of targeted improvements to MonoDGP for more robust monocular 3D object detection, with an emphasis on stabilizing and selectively trusting depth guidance. First, we treat projection-based geometric depth as a detached prior and supervise a learnable geometry-error residual, which stabilizes optimization and

Fig. 4 Qualitative results on the KITTI validation set are presented. The left column shows the detection results on the original images, while the second and third columns provide zoomed-in comparison results focusing on VRU detections. Ground-truth boxes are color-coded based on their class: Pedestrian (green), Car (blue), and Cyclist (orange). Predicted boxes are also class-specific, with Pedestrian boxes in yellow, Car boxes in light pink, and Cyclist boxes in red

improves robustness for distant objects and VRUs. Second, we decouple the 2D and 3D prediction heads to reduce interference between image-plane perception and 3D geometric reasoning, providing a better structural basis for subsequent depth-aware fusion. Third, we introduce query-adaptive gated depth cross-attention in the 3D decoder to suppress unreliable depth cues on a per-query basis and better exploit reliable depth evidence when available. In addition, we enhance the region segmentation module with learnable multi-scale fusion, channel-spatial attention, and residual gating to produce more stable region-conditioned representations. Ablation studies show that the proposed components play different but complementary roles: DDP provides stable geometric supervision, ADG is the main driver of depth-aware performance gain, E-RSM improves region-conditioned feature quality, and DPH contributes as a structural decoupling design that is most effective when combined with the later depth-aware modules. Experiments on KITTI demonstrate overall improvements over MonoDGP, with particularly strong gains on Pedestrian and competitive results on Car and Cyclist.

Acknowledgement

This work was supported by the Ministry of Trade, Industry & Energy (MOTIE) and the Korea Evaluation Institute of Industrial Technology (KEIT) through the project "Infrastructure for 3P service providers to develop, test, validate & operate services on the new controller" [Project Number: 2410012549, RS-2024-00506825].

References

1) C. R. Qi, H. Su, K. Mo and L. C. R. Qi, H. Su, K. Mo and L. J. Guibas, J. Guibas, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.652-660, 2017.
2) T. Yin, X. Zhou and P. Krahenbuhl, "Center-Based 3D Object Detection and Tracking," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.11784-11793, 2021.
3) S. Shi, L. Jiang, J. Deng, Z. Wang, C. Guo, J. Shi, X. Wang and H. Li, "PV-RCNN++: Point-Voxel Feature Set Abstraction with Local Vector Representation for 3D Object Detection," International Journal of Computer Vision, Vol.131, No.2, pp.531-551, 2023.
4) Y. Chen, J. Liu, X. Zhang, X. Qi and J. Jia, "VoxelNeXt: Fully Sparse VoxelNet for 3D Object Detection and Tracking," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.21674-21683, 2023.
5) T. Peng and B. Kim, "Improving Accuracy of Pseudo-LiDAR for 3D Object Detection by Accurate Depth Estimation," 2023 IEEE 6th International Conference on Knowledge Innovation and Invention (ICKII), IEEE, 2023.
6) P. Liao, F. Yang, D. Wu, W. Zhao and J. Yu, "MonoDETRNext: Next-Generation Accurate and Efficient Monocular 3D Object Detector," arXiv Preprint arXiv:2405.15176, 2024.
7) T. Wang, X. Zhu, J. Pang and D. Lin, "FCOS3D: Fully Convolutional One-Stage Monocular 3D Object Detection," Proceedings of the IEEE/CVF International Conference on Computer Vision, pp.913-922, 2021.
8) S. Luo, H. Dai, L. Shao and Y. Ding, "M3DSSD: Monocular 3D Single Stage Object Detector," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.6141-6150, 2021.
9) K. C. Huang, T. H. Wu, H. T. Su and W. H. Hsu, "MonoDTR: Monocular 3D Object Detection with Depth-Aware Transformer," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.4012-4021, 2022.
10) C. Pan, J. Peng and Z. Zhang, "Depth-Guided Vision Transformer with Normalizing Flows for Monocular 3D Object Detection," IEEE/CAA Journal of Automatica Sinica, Vol.11, No.3, pp.673-689, 2024.
11) T. Peng, J. An and B. Kim, "Dynamic Feature Fusion for Depth-Guided Transformer in Monocular 3D Object Detection by Adaptive BiFPN," KSAE Spring Conference Proceedings, pp.1514-1519, 2025.
12) 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," Proceedings of the IEEE/CVF International Conference on Computer Vision, pp.9121-9132, 2023.
13) F. Pu, Y. Wang, J. Deng and W. Yang, "MonoDGP: Monocular 3D Object Detection with Decoupled-Query and Geometry-Error Priors," Proceedings of the Computer Vision and Pattern Recognition Conference, pp.6520-6530, 2025.
14) 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.

15) N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov and S. Zagoruyko, "End-to-End Object Detection with Transformers," European Conference on Computer Vision, Springer International Publishing, Cham, pp.213-229, 2020.
16) X. Chen, K. Kundu, Y. Zhu, A. G. Berneshawi, H. Ma, S. Fidler and R. Urtasun, "3D Object Proposals for Accurate Object Class Detection," Advances in Neural Information Processing Systems, Vol.28, 2015.
17) A. Simonelli, S. R. Bulo, L. Porzi, M. Lopez-Antequera and P. Kontschieder, "Disentangling Monocular 3D Object Detection," Proceedings of the IEEE/CVF International Conference on Computer Vision, pp.1991-1999, 2019.
18) H. Chen, Y. Huang, W. Tian, Z. Gao and L. Xiong, "MonoRUn: Monocular 3D Object Detection by Reconstruction and Uncertainty Propagation," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.
19) C. Reading, A. Harakeh, J. Chae and S. L. Waslander, "Categorical Depth Distribution Network for Monocular 3D Object Detection," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.8555-8564, 2021.
20) M. Ding, Y. Huo, H. Yi, Z. Wang, J. Shi, Z. Lu and P. Luo, "Learning Depth-Guided Convolutions for Monocular 3D Object Detection," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp.1000-1001, 2020.
21) 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," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.454-463, 2021.
22) Y. Chen, L. Tai, K. Sun and M. Li, "MonoPair: Monocular 3D Object Detection Using Pairwise Spatial Relationships," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.12093-12102, 2020.
23) A. Kumar, G. Brazil, E. Corona, A. Parchami and X. Liu, "DEVIANT: Depth Equivariant Network for Monocular 3D Object Detection," European Conference on Computer Vision, Springer Nature Switzerland, Cham, pp.664-683, 2022.
24) S. Luo, H. Dai, L. Shao and Y. Ding, "M3DSSD: Monocular 3D Single Stage Object Detector," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.6145-6154, 2021.
25) X. Ma, Y. Zhang, D. Xu, D. Zhou, S. Yi, H. Li and W. Ouyang, "Delving into Localization Errors for Monocular 3D Object Detection," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.4721-4730, 2021.
26) Y. Zhang, J. Lu and J. Zhou, "Objects Are Different: Flexible Monocular 3D Object Detection," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.
27) L. Peng, X. Wu, Z. Yang, H. Liu and D. Cai, "DID-M3D: Decoupling Instance Depth for Monocular 3D Object Detection," European Conference on Computer Vision, Springer Nature Switzerland, Cham, pp.1-18, 2022.
28) Y. Lu, X. Ma, L. Yang, T. Zhang, Y. Liu, Q. Chu, J. Yan and W. Ouyang, "Geometry Uncertainty Projection Network for Monocular 3D Object Detection," Proceedings of the IEEE/CVF International Conference on Computer Vision, pp.3111-3121, 2021.