# A Visual Context Enhancement Module for Transformer-based Monocular 3D Object Detection

Tao Peng*, Jinsu An*, Byeong Woo Kim*
*Dept. of Electrical, Electronic and Computer Engineering, University of Ulsan
e-mail: bywokim@ulsan.ac.kr

# 트랜스포머 기반 단안 3D 객체 검출을 위한 시각적 컨텍스트 향상 모듈

팽 도*, 안진수*, 김병우*
*울산대학교 전기전자컴퓨터공학과

Abstract

Monocular 3D object detection is highly demanded in various industrial applications, including robotics, smart surveillance, and automated guided vehicles (AGVs), due to its cost-effectiveness and versatile deployment compared to expensive LiDAR systems. However, existing Transformer-based methods struggle to accurately localize distant, small, or heavily occluded objects, limiting their practical deployment in complex real-world environments. To address this, we propose a lightweight Visual Context Enhancement Module (VCEM). VCEM utilizes a multi-branch convolutional mechanism and adaptive channel attention to efficiently expand the receptive field and highlight critical object boundaries. As a plug-and-play component, VCEM seamlessly integrates into existing baselines with minimal computational overhead. Our approach significantly improves detection accuracy for challenging instances, demonstrating strong potential for cost-effective 3D perception in broad academia-industry applications.

## 1. Introduction

The demand for 3D environmental perception has surged across various industrial applications, including smart factories, Advanced Driver Assistance Systems (ADAS), and intelligent surveillance robots. While LiDAR sensors have traditionally been the standard for 3D perception, their high deployment costs and complex maintenance requirements severely hinder widespread industrial adoption. In contrast, monocular 3D object detection, which relies entirely on a single camera [1, 2, 3], offers a highly cost-effective and easily deployable alternative.

Recently, Transformer-based monocular 3D object detection architectures [1, 5] have achieved remarkable performance improvements. However, in complex real-world environments, these methods struggle to accurately localize distant, occluded, or small objects.

To address this limitation, we propose a lightweight Visual Context Enhancement Module (VCEM) designed to maximize the performance of Transformer-based monocular 3D object detection frameworks. As a plug-and-play component, VCEM is strategically positioned as an input-side token enhancer prior to the Transformer encoder to deeply enrich visual representations.

The main contributions of this paper are summarized as follows:

- We analyze the causes of performance degradation in monocular 3D perception within complex industrial environments and propose VCEM, a plug-and-play module tailored for cost-effective practical deployment.
- Extensive experiments on the challenging KITTI 3D benchmark demonstrate that our proposed method achieves significant performance gains on "Hard" difficulty instances while maintaining inference efficiency.

## 2. Related Work

### 2.1 3D Object Detection

Accurate 3D object detection is a foundational technology for various industrial applications. Currently, LiDAR-based methods[4] utilize precise point cloud data to achieve exceptional localization accuracy. However, the high cost, mechanical fragility, and computational demands of LiDAR sensors heavily restrict their deployment in cost-sensitive industrial applications, such as ADAS. While stereo and multi-camera systems[6] provide explicit geometric cues, they require rigorous hardware synchronization and baseline calibration. Consequently, monocular 3D object detection—relying entirely on a single standard camera—has drawn massive attention due to its unparalleled deployment flexibility and minimal hardware cost.

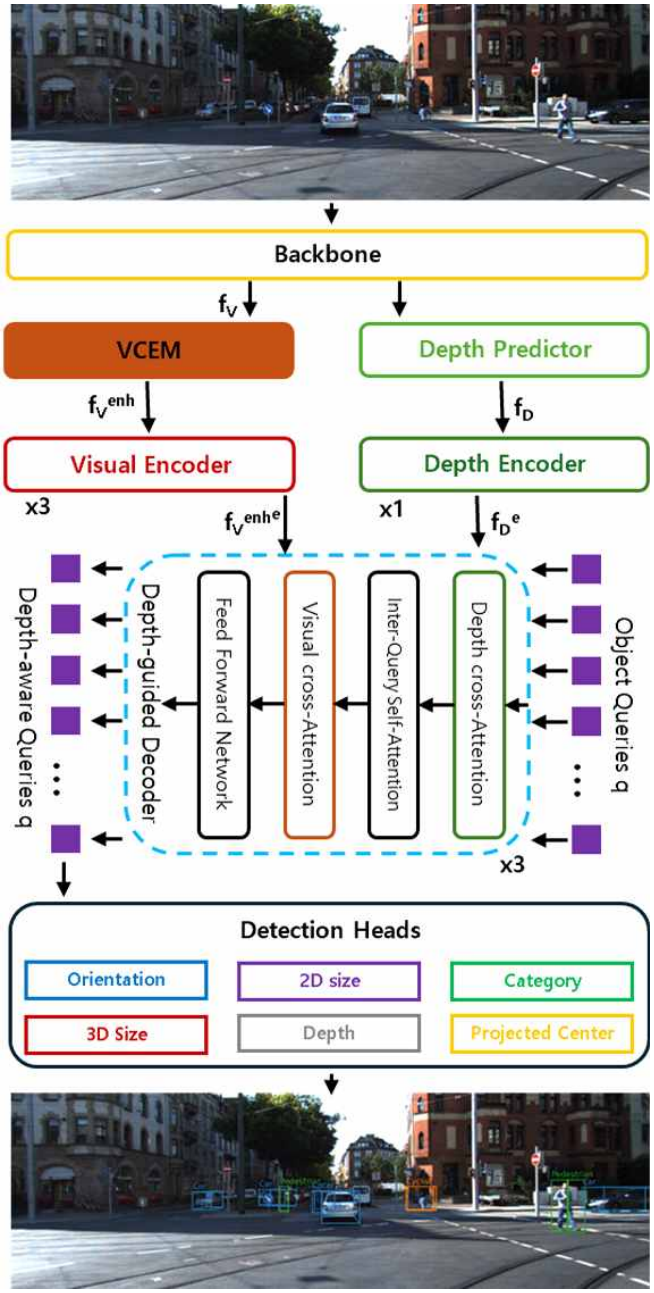
### 2.2 Monocular 3D Object Detection

Estimating 3D bounding boxes from a single 2D image is an inherently ill-posed problem. Early CNN-based frameworks[7] depended on geometric priors and 2D-to-3D projection constraints but frequently encountered depth ambiguity. Recently, Transformer-based architectures, such as MonoDETR[5], successfully connect visual features and 3D space by using depth-guided attention mechanisms. Despite these advancements, existing visual encoders in these frameworks still extract high-resolution, low-level features with restricted receptive fields. This leads to suboptimal context modeling, making it difficult to localize distant, small, or severely occluded targets. While prior works focus heavily on refining the depth predictor, our approach uniquely targets the visual branch. We introduce a lightweight enhancement module to explicitly expand the visual receptive field and recalibrate feature channels before the decoding stage, ensuring robust perception for challenging instances.

## 3. Methodology

### 3.1 Overall Architecture

As illustrated in Fig. 1, our framework builds upon MonoDETR. While the baseline extracts multi-scale visual features for Transformer-based 3D regression, its high-resolution features—crucial for small and distant objects—suffer from restricted receptive fields. To address this, we insert the lightweight, plug-and-play VCEM at the crucial juncture between the visual projection layer and the depth-aware Transformer, deeply enriching the features' contextual and geometric representations before global self-attention occurs.




[Fig. 1] Overall architecture of the proposed method. The lightweight VCEM operates as an input-side token enhancer, enriching visual features before they enter the depth-aware Transformer.

### 3.2 Visual Context Enhancement Module

The core objective of VCEM is to expand the receptive field to capture surrounding context while explicitly highlighting object boundaries. Let $F_v \in R^{C\times H\times W}$ denote the high-resolution visual feature map outputed by the visual encoder. The enhanced visual feature $F_v^{enh}$ is formulated as:

$$F_v^{enh} = F_v + \alpha \cdot ECA(MSConv(F_v)) \quad (1)$$

where $MSConv(\cdot)$ denoted the Multi-branch Convolution mechanism, $ECA(\cdot)$ represents the Efficient Channel Attention, and $\alpha$ is a learnable scaling coefficient. The module operates through two key steps:

• Multi-Scale Context Expansion(MSConv): To accurately localize distant and occluded targets typical in real-world deployments , a broad receptive field is mandatory. MSConv utilizes parallel branches with varying dilation rates to aggregate multi-scale spatial context. This allows the network to explicitly capture the broader surroundings of an object, effectively overcoming the strict local limitations of standard convolutions.

• Adaptive Feature Recalibration(ECA): Once the spatial context is enriched, it is vital to filter out background noise and emphasize discriminative local textures and depth edges. The ECA block adaptively recalibrates the channel-wise weights of the context-enriched features using a fast 1D convolution. This step ensures that the network selectively focuses on critical geometric boundaries.

Finally, a residual connection is employed, scaled by the learnable parameter $\alpha$. This residual design ensures stable gradient flow during early training stages and guarantees that the VCEM acts as a strictly additive enhancement, preserving the required real-time inference efficiency.

## 4. Experiments

### 4.1 Experimental Setup

We evaluate our proposed method on the challenging KITTI 3D object detection benchmark, utilizing the standard split of 3,712 training and 3,769 validation samples. All experiments are built upon the MonoDETR baseline with a ResNet-50 backbone. For training, we employ the AdamW optimizer with a batch size of 16 for a total of 195 epochs.

### 4.2 Main Results

As shown in Table 1, the integration of our proposed VCEM demonstrates a clear advantage over the baseline MonoDETR, particularly in more challenging perception scenarios. While maintaining competitive performance on the "Easy" subset, our method achieves notable

improvements on the "Moderate" and "Hard" difficulties. Furthermore, our method provides robust Bird's Eye View localization.

[Table 1] Quantitative comparison with the baseline MonoDETR on the KITTI validation set.

| Method | Car, $AP_{3D}$ / $AP_{BEV}$, Iou=0.7 | | |
|---|---|---|---|
| | Easy | Mod. | Hard |
| MonoDETR | **28.84** | 20.61 | 16.38 |
| Ours | 28.41 | **20.92** | **17.48** |

## 4.3 Ablation Study

To verify the contribution of each component within VCEM, we perform a detailed ablation study, as summarized in Table 2.

• Baseline + MSConv: Expanding the receptive field via multi-scale convolutions leads to a steady increase in detection accuracy for distant objects.

• Baseline + MSConv + ECA (Full VCEM): Adding the channel attention mechanism provides the best results, as it allows the network to suppress background noise and focus on critical geometric edges, validating the synergy between spatial context and feature recalibration.

[Table 2] Ablation study of the proposed components in VCEM on the KITTI validation set.

| Method | Car, $AP_{3D}$ / $AP_{BEV}$, Iou=0.7 | | |
|---|---|---|---|
| | Easy | Mod. | Hard |
| Baseline | 26.69/36.15 | 19.68/26.25 | 15.87/22.36 |
| + MSConv | 27.32/36.78 | 19.52/26.50 | 16.30/22.70 |
| + MSConv + ECA | 28.41/38.38 | 20.92/27.24 | 17.48/23.20 |

## 4.4 Qualitative Results

Fig. 2 provides a visual comparison between the baseline and our proposed method. While the baseline frequently misses small-scale objects at long distances or fails to localize objects in cluttered scenes, our VCEM-enhanced model successfully recovers these targets with precise 3D bounding boxes. This qualitative evidence confirms that richer visual context and geometric awareness lead to more robust perception in complex real-world settings.


[Fig. 1] Qualitative landmark prediction visualization over MonoDETR.

# 5. Conclusion

In this paper, we presented the VCEM to address the limitations of monocular 3D object detection in complex industrial environments. By synergistically combining multi-scale context expansion and adaptive feature recalibration, VCEM effectively broadens the visual receptive field and highlights critical geometric boundaries. Experiments on the KITTI benchmark demonstrate that our plug-and-play module significantly boosts detection accuracy for distant and highly occluded targets without compromising real-time inference efficiency. Ultimately, this work provides a robust and cost-effective 3D perception solution, showing great potential for large scale deployment in ADAS and smart factory systems.

## 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] Yan, Longfei, et al. "Monocd: Monocular 3d object detection with complementary depths." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

[2] Peng, Tao, and Byeongwoo 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.

[3] Peng, Tao, Jinsu An, and ByeongWoo Kim. "Dynamic Feature Fusion for Depth-Guided Transformer in Monocular 3D Object Detection by Adaptive BiFPN." 한국자동차공학회 춘계학술대회 (2025): 1514-1519.

[4] Aung, Nang Htet Htet, et al. "A review of lidar-based 3d object detection via deep learning approaches towards robust connected and autonomous vehicles." IEEE Transactions on Intelligent Vehicles (2024).

[5] Zhang, Renrui, et al. "Monodetr: Depth-guided transformer for monocular 3d object detection." Proceedings of the IEEE/CVF international conference on computer vision. 2023.

[6] Mu, Shiyi, et al. "StereoDETR: Stereo-based Transformer for 3D Object Detection." IEEE Transactions on Circuits and Systems for Video Technology (2025).

[7] Chen, Xiaozhi, et al. "Monocular 3d object detection for autonomous driving." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.