# Monocular Vehicle Pose and Shape Reconstruction via Dynamic Context Adaptation and Progressive Geometry Refinement

**Wei Li, Long Ji, Ying Wang, Xiao Wu, Zhaoquan Yuan*, Penglin Dai**

School of Computing and Artificial Intelligence, Southwest Jiaotong University
Engineering Research Center of Sustainable Urban Intelligent Transportation, Ministry of Education
{liwei,zqyuan,wuxiaohk,penglindai}@swjtu.edu.cn, jl1207@my.swjtu.edu.cn, wangying920506@gmail.com

## Abstract

Accurate reconstruction of 3D vehicle pose and shape from monocular images is challenging, particularly for distant objects in autonomous driving. Existing methods often suffer from geometric ambiguity in depth estimation and structural hollowness in shape recovery, primarily due to inadequate multi-scale feature aggregation and unflexible prior modeling. To overcome these limitations, MonoVPR is proposed, a novel framework integrating dynamic context adaptation and progressive geometry refinement. Specifically, a Hierarchical Dual-Context Attention (HDCA) module is introduced to resolve scale-dependent degradation through gated cross-attention across multi-resolution feature maps, dynamically fusing object-centric geometric cues with scene-centric semantics. For shape refinement, the Bounded Iterative Mesh Refiner (BIMR) progressively optimizes template-guided deformations via multi-head attention and a tanh-bounded correction loop, ensuring physically plausible reconstructions. Extensive experiments on the ApolloCar3D benchmark demonstrate MonoVPR achieves state-of-the-art performance, showing exceptional capability in reconstructing geometrically consistent shapes and precise poses for challenging long-range scenarios.

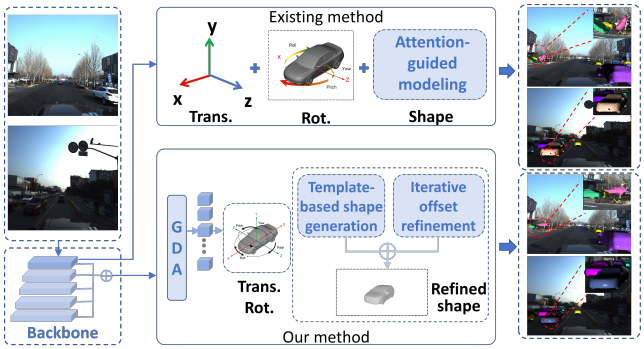


Figure 1: The existing method often struggles with distant objects. Its limitations stem from two principal shortcomings: first, the failure to effectively integrate multi-scale visual cues undermines the robustness of pose estimation; second, the inflexible single-pass deformation strategy leads to geometrically implausible shape artifacts.. These challenges motivate our work, MonoVPR, which is designed to tackle them through dynamic context adaptation and progressive geometry refinement.

## Introduction

Three-dimensional vehicle reconstruction (estimating 3D pose and shape) is a fundamental task in computer vision, crucial for applications like autonomous driving by providing perception for safe navigation, motion planning, and comprehensive scene understanding. Although high-precision 3D perception is successfully demonstrated using active sensors such as LiDAR (Bai et al. 2022; Erçelik et al. 2022; Shi et al. 2020) or stereo camera systems (Li, Chen, and Shen 2019; Sun et al. 2020), these solutions have drawbacks: high hardware costs, operational limitations (e.g., sparse point clouds at long ranges), and complex calibration requirements. Consequently, the development of robust, low-cost monocular methods that infer 3D structure from a single RGB image is becoming an emerging and critical research focus (Lu et al. 2021; Zhang, Lu, and Zhou 2021).

Recovering 3D information from a single 2D projection is an inherently ill-posed inverse problem. A single image lacks explicit depth cues for unambiguous 3D location, which means that any given 2D projection could correspond to infinite possible 3D scenes (Liu et al. 2021). MonoFlex (Zhang, Lu, and Zhou 2021) focuses on geometry constraints between 2D and 3D. GSNet (Ke et al. 2020) designs a PCA-based model for vehicle shape reconstruction. Recently, BAAM (Lee et al. 2023) uses a bi-contextual attention module with an attention-guided modeling mechanism, achieving notable performance by explicitly integrating inter-object relations and scene-level contextual cues. However, current monocular 3D reconstruction methods exhibit severe failure modes in challenging real-world settings, struggling with cluttered scenes, significant articulations or viewpoint changes, and distant objects that provide limited visual cues.

As shown in Figure 1, the existing method typically extracts features and processes them through parallel branches, employing separate streams that rely primarily on attention-guided modeling to capture global context information. Nevertheless, the absence of an effective strategy for integrating

*Corresponding author.

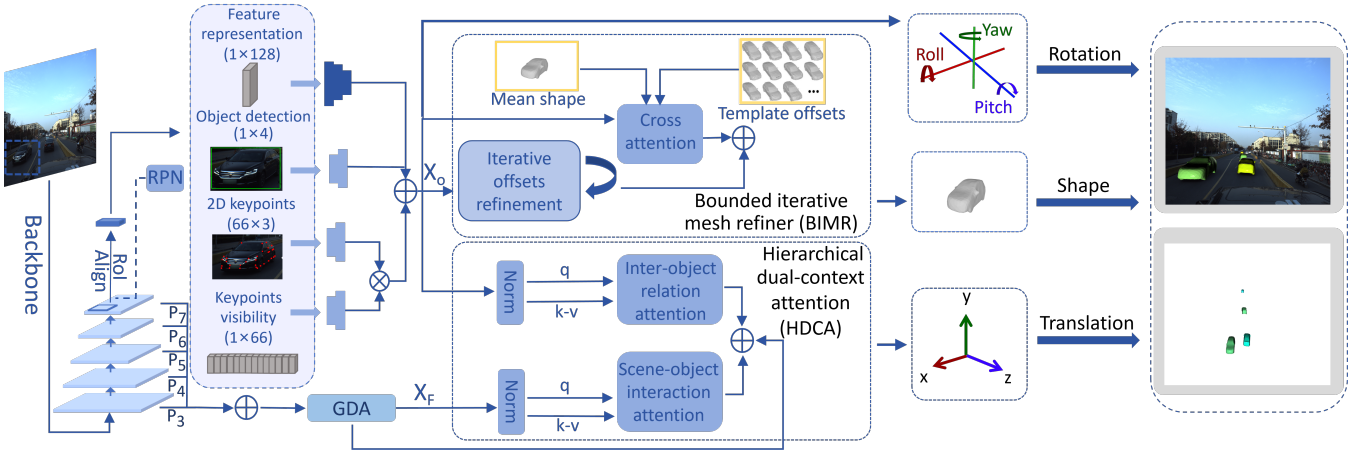


Figure 2: An overview of the proposed MonoVPR framework. MonoVPR first extracts an object-centric feature from 2D primitives and a hierarchical scene context from the backbone's feature pyramid using a GDA block. The framework then processes this information through two specialized streams. HDCA fuses the object-centric and scene context features through parallel inter-object and scene-object attention mechanisms to produce an enhanced representation for robust pose estimation. Concurrently, BIMR progressively optimizes a 3D mesh by combining template-guided deformations with an iterative refinement of vertex offsets. This dual-stream architecture enables the precise reconstruction of both 3D pose and shape for challenging vehicle instances.

fine-grained local features restricts information flow across scales, which in turn limits pose estimation accuracy. Moreover, recovering shape remains particularly difficult, as directly regressing mesh vertex coordinates from image features is susceptible to prediction errors. When such inaccuracies are propagated in a single-step template deformation using a shape prior model, they often introduce severe structural artifacts and compromise geometric consistency (Lee et al. 2023). This inflexible paradigm fails to enable localized corrections, frequently resulting in hollow and unrealistic reconstructions.

To address these challenges, this paper presents MonoVPR, a novel framework that redefines the reconstruction pipeline through two dedicated components, as illustrated in Figure 2. First, the Hierarchical Dual-Context Attention (HDCA) module is designed to overcome limitations in multi-scale feature integration. By employing gated cross-attention across a feature pyramid, it enables dynamic integration of object-centric geometric cues and scene-centric semantic information, substantially improving pose estimation accuracy. Second, to overcome the limitations of single-step shape deformation, the Bounded Iterative Mesh Refiner (BIMR) is developed. Built on a learnable template, BIMR adopts a progressive refinement approach instead of predicting a single imprecise transformation. It integrates multi-head attention with a tanh-bounded correction loop to implement a series of localized vertex adjustments. This iterative optimization enhances topological inconsistencies, eliminates structural hollowness and restores geometric consistency in the reconstructed shape. The experimental results on the large-scale ApolloCar3D benchmark show that MonoVPR significantly outperforms current state-of-the-art methods in multiple key metrics. The main contributions can be summarized as follows:

- A novel end-to-end framework, termed MonoVPR, is proposed to address the challenges of geometric ambiguity and structural hollowness by integrating dynamic context adaptation with iterative geometry refinement.
- A Hierarchical Dual-Context Attention (HDCA) module is designed to enhance scale-invariant features through the dynamic fusion of object-level geometry and scene-level semantics from multi-resolution maps.
- A Bounded Iterative Mesh Refiner (BIMR) module is introduced to achieve physically plausible shapes by replacing single-step deformation with an iterative correction mechanism that constrains vertex displacements.

## Related Work

### Monocular 3D Pose and Shape Reconstruction

Early methods largely depend on strong geometric priors. DeepMANTA (Chabot et al. 2017) performs matching against CAD libraries, while 3D-RCNN (Kundu, Li, and Rehg 2018) leverages PCA-based shape subspaces. Mono3D++ (He and Soatto 2019) further incorporates task-specific constraints through an optimization framework. Alternative strategies introduce intermediate representations, such as the fusion of RGB and depth features in RoI-10D (Manhardt, Kehl, and Gaidon 2019). Recently, context-aware joint reasoning is adopted: GSNet (Ke et al. 2020) introduces multi-way feature fusion, BAAM (Lee et al. 2023) explores bi-contextual attention, and DAGM-Mono (Murhij and Yudin 2024) utilizes deformable attention mechanisms. Another line of work formulates the task as a generative problem (Zhang et al. 2021). MonoDiff (Ranasinghe, Hegde, and Patel 2024) treats pose estimation as a reverse diffusion process where parameters are iteratively refined from noise conditioned on 2D image features. In contrast,

ZeroShape (Huang et al. 2024) reasserts the effectiveness of regression-based approaches by jointly estimating depth and camera intrinsics to form accurate visible surface representations, an idea extended by ZeroShape-W (Cho et al. 2025) to handle in-the-wild images through explicit occluder mask regression. Separately, Mono3R (Li et al. 2025) applies robust monocular geometric priors to enhance multi-view reconstruction, particularly in texture-scarce regions. Despite these advances, existing methods remain challenged by insufficient aggregation of multi-scale features and inflexible shape prior modeling, resulting in persistent geometric ambiguity and structurally incomplete shape recovery.

### Monocular 3D Object Detection

Monocular 3D object detection aims to infer 3D properties from a single image. Depth-assisted approaches, such as Pseudo-LiDAR (Wang et al. 2019) and DD3D (Park et al. 2021), rely on intermediate depth representations. OccupancyM3D (Peng et al. 2024) adopts a related strategy by incorporating sparse LiDAR supervision during training to enhance voxel-based occupancy learning. Other methods, including MonoCD (Yan et al. 2024) and DP-M3D (Shi et al. 2025), explore specialized transformer architectures to improve the fusion of depth and image features. To address fundamental geometric inaccuracies, MonoDGP (Pu et al. 2025) introduces a geometry error prior for correcting standard perspective projection. In Bird's-Eye-View (BEV) based techniques, TaDe (Zhao et al. 2024) decomposes the perspective-to-BEV transformation into two distinct stages, while UniMODE (Li et al. 2024) employs an uneven BEV grid with increased resolution near the camera to handle diverse object scales. ADD (Wu et al. 2023), applies knowledge distillation to transfer geometric awareness from a depth-aware teacher model without increasing inference cost. Alternatively, generative formulations such as MonoDiff (Ranasinghe, Hegde, and Patel 2024) treat 3D detection as a reverse diffusion process, refining bounding box parameters from 2D features using a Gaussian Mixture Model. Monocular 3D pose and shape reconstruction provides a more detailed geometric understanding of object structure, complementing and enriching 3D bounding box estimation.

## Methods

### Pipeline

The proposed MonoVPR framework is built upon a standard 2D object detector, such as Mask R-CNN (He et al. 2017). Initially, a hierarchical feature pyramid is generated from an input RGB image using a Res2Net backbone (Gao et al. 2019) integrated with BiFPN (Tan, Pang, and Le 2020). Essential 2D primitives, including object detection, 2D keypoints, and their visibility scores, are then extracted for each vehicle instance. These primitives are processed and unified into a object-centric feature map, denoted as $X_o$, which encodes the distinctive appearance and 2D spatial attributes of each detected object.

The 3D vehicle representation is decomposed into parallel streams for translation, rotation, and shape. The Hierarchical Dual-Context Attention (HDCA) module is employed for robust 3D translation estimation, where a comprehensive scene context is formulated from the pyramidal features and seamlessly fused with the object-centric features. For 3D rotation, the parameters are directly regressed from the object-centric feature vector through a sequence of three fully connected layers. In parallel, the Bounded Iterative Mesh Refiner (BIMR) module is dedicated to shape recovery, where a template mesh is first employed and then its vertex positions are updated through a bounded iterative process to accurately capture fine-grained geometric details. Finally, a precise 6-DoF pose alongside a geometrically coherent 3D mesh is obtained for each vehicle instance.

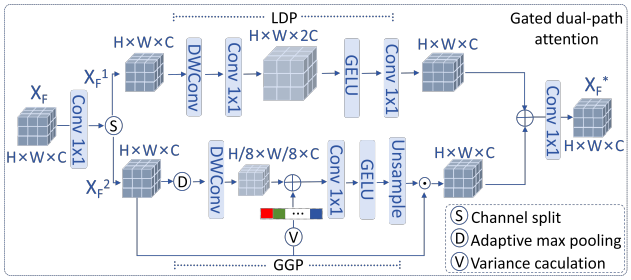


Figure 3: The architecture of GDA, designed to generate an enhanced scene-centric representation by processing features through two complementary pathways. LDP is responsible for capturing fine-grained spatial patterns, while GGP extracts scale-invariant global statistics to emphasize salient features.

### Hierarchical Dual-Context Attention

Estimating 3D translation from a single image is inherently ambiguous due to the depth-scale uncertainty. Existing approaches typically depend on spatially coarse feature maps, which prove insufficient as they fail to preserve fine-grained local information crucial for precise vehicle placement (Peng et al. 2024). To address this issue, the Hierarchical Dual-Context Attention (HDCA) module is introduced, which bridges this gap by formulating a multi-scale scene context and integrating it with object-specific features. This synergistic fusion produces a context-aware representation that significantly mitigates depth ambiguity for robust translation estimation.

The hierarchical context formulation begins by integrating global semantic information from multiple levels of the backbone's feature pyramid. To ensure dimensional consistency, higher-resolution feature maps are spatially downsampled to a uniform resolution via bilinear interpolation. These aligned multi-scale features are then concatenated along the channel dimension, synthesizing a comprehensive feature representation that spans from local details to global abstractions. A $1\times1$ convolutional layer is subsequently applied to this concatenated representation to reduce channel redundancy, resulting in fused feature map, denoted as $X_F$. The feature map is then enhanced by the Gated Dual-path Attention (GDA) block. To capture complementary information, $X_F$ is first transformed and flow into two parallel

streams $X_F^1$ and $X_F^2$. The feature map of the first stream $X_F^1$, processed by the Local Detail Path (LDP) through a convolutional block, is focused on capturing fine-grained spatial patterns, i.e., $X_{\text{LDP}} = \text{LDP}(X_F^1)$.

In parallel, the feature stream $X_F^2$ is modulated by the Global Gated Path (GGP). A dynamic gating signal $G_{\text{gate}}$ is generated through the integration of two global statistical descriptors: a spatial summary derived via max pooling ($x_s$) and the channel-wise variance ($x_v$) of the input features. This signal is subsequently upsampled to match the original feature dimensions and applied through element-wise multiplication to the feature stream $X_F^2$, thereby selectively emphasizing the most informative regions. The gating signal is formally defined as:

$$G_{\text{gate}} = \mathcal{I}_{\text{up}}(\text{GELU}(\text{Conv}_{1\times1}(\alpha \odot x_s + \beta \odot x_v))), \quad (1)$$

where $\mathcal{I}_{\text{up}}$ denotes an upsampling interpolation operator that restores the spatial resolution. GELU represents the Gaussian Error Linear Unit activation function. The scalars $\alpha$ and $\beta$ are learnable parameters introduced to apply a bounded affine transformation to the input feature stream, providing a stable and adaptable scaling mechanism. The feature representation $X_{\text{GGP}}$ of GGP is obtained by applying the dynamic gating signal to the input feature stream $X_F^2$ via element-wise multiplication, which can be written as:

$$X_{\text{GGP}} = X_F^2 \odot G_{\text{gate}}. \quad (2)$$

The complementary feature representations generated by the two parallel paths are first aggregated through summation. This combined feature is then processed by a final $1\times1$ convolutional layer to produce a refined contextual representation $X_F^* = \text{Conv}_{1\times1}(X_{\text{LDP}} + X_{\text{GGP}})$.

Following $X_F^*$, dual-context attentive fusion is performed through two complementary attention mechanisms. First, the inter-object relation attention is applied, where spatial dependencies among vehicle instances are modeled via multi-head self-attention, yielding a relation-aware feature $X_{\text{rel}}$. Second, the scene-object interaction attention is employed, utilizing a cross-attention mechanism to situate each object within its environmental context. In this process, object features serve as queries while the global context provides keys and values, producing a context-aware feature $X_{\text{ctx}}$. These two enriched features are integrated with the original object features $X_o$ via residual connections, with channel-wise adaptive weighting performed by learnable scaling matrices $\lambda_{\text{rel}}$ and $\lambda_{\text{ctx}}$, yielding the final enhanced representation $\tilde{X}_t$:

$$\tilde{X}_t = X_o + \lambda_{\text{rel}} \odot X_{\text{rel}} + \lambda_{\text{ctx}} \odot X_{\text{ctx}}, \quad (3)$$

where the original object features are augmented by the weighted relation-aware and context-aware features.

## Bounded Iterative Mesh Refiner

Detailed 3D shape reconstruction remains difficult, as conventional single-step deformation approaches often produce erroneous vertex offsets that lead to distorted and topologically unsound meshes with structural hollowness.The proposed Bounded Iterative Mesh Refiner (BIMR) addresses this by progressively optimizing an initial template shape. A core component of this module is a tanh-bounded correction loop that applies a series of conditioned, stable vertex updates. The risk of drastic geometric artifacts is thus prevented, yielding physically plausible and coherent 3D vehicle models.

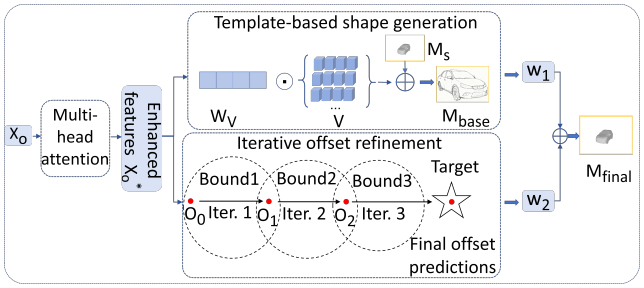


Figure 4: The architecture of BIMR, which produces a high-fidelity mesh using enhanced object-centric features via two collaborative components. In parallel, the template-Based shape generation component constructs a base mesh from deformable priors, while the Iterative Offset Refinement component applies bounded corrections to capture geometric details.

BIMR is initiated by refining the object-centric features, followed by the execution of two parallel shape prediction pathways, and finalized through the integration of their results. As depicted in Figure 4, the object-centric features $X_o$ are first processed by a series of Multi-Head Attention (MHA) layers for enhancement. At each layer $l$, the object feature $X_o^{(l)}$ is employed as a query to interact with a set of learnable shape-concept embeddings $E$, thereby generating an improved feature $X_o^{(l+1)}$ for subsequent processing:

$$X_o^{(l+1)} = \text{LayerNorm}(X_o^{(l)} + \text{MHA}(Q = X_o^{(l)}, K = E, V = E)). \quad (4)$$

The resulting enriched feature $X_o^*$ drives two concurrent shape estimation branches. In the first branch, template-based shape generation is performed. Here, $X_o^*$ is mapped to a set of weighting coefficients $W_V$ over a deformable shape basis $V$. These coefficients are used to deform a canonical template $M_s$, yielding a base mesh $M_{\text{base}}$:

$$W_V = \text{Softmax}(\text{Linear}(X_o^*)), \quad (5)$$

$$M_{\text{base}} = M_s + (W_V \cdot V). \quad (6)$$

Simultaneously, the second branch conducts iterative offset refinement to recover instance-specific geometry. A coarse initial offset $O^{(0)}$ is first generated from $X_o^*$, then refined over multiple steps. At iteration $t$, an adjustment term $\Delta O^{(t)}$ is predicted by a shared MLP block that incorporates both the static feature $X_o^*$ and the preceding offset $O^{(t-1)}$. A bounded update rule is applied to ensure numerical stability:

$$O^{(0)} = \text{MLP}(X_o^*), \quad (7)$$

$$\Delta O^{(t)} = \text{MLP}([X_o^*; O^{(t-1)}]), \quad (8)$$

$$O^{(t)} = O^{(t-1)} + 0.5 \cdot \tanh(\Delta O^{(t)}). \quad (9)$$

The hyperbolic tangent activation $\tanh$ confines each coordinate update to the range $(-0.5, 0.5)$, preventing unrealistic shape deviations. The outputs from both branches are combined via learnable scalars $w_1$ and $w_2$ to produce the final reconstructed mesh $M_{\text{final}}$:

$$M_{\text{final}} = w_1 M_{\text{base}} + w_2 O^{(T)}, \quad (10)$$

where $T$ refers to the number of iterations. Through this dual-branch architecture, the module effectively combines template-driven shape priors with instance-wise geometric adjustments, enabling high-fidelity and topologically sound mesh reconstruction.

## Loss Function

The loss formulation in MonoVPR follows BAAM (Lee et al. 2023).The network is trained end-to-end using a multi-task loss function that encompasses regression losses for 3D parameters, standard detection losses, and a 3D spatial loss to ensure geometric consistency. Specifically, the regression losses include translation loss $L_t$, rotation loss $L_r$, and shape loss $L_{\text{shape}}$. The translation loss uses an $L_1$ norm for XY-plane coordinates and an uncertainty-aware $L_1$ norm for depth. The rotation loss is a cyclical $L_1$ loss handling angle periodicity. The shape loss is the mean $L_2$ distance between the predicted and ground-truth mesh vertices. The detection loss supervises the 2D object detection pretext task and is a standard combination of losses from the underlying Mask R-CNN (He et al. 2017) framework. Furthermore, the 3D spatial loss is employed to account for the crucial interdependence of translation, rotation, and shape. This loss penalizes the mean vertex error between predicted and ground-truth meshes across three disentangled coordinate spaces: a rotation-only space, a translation-only space, and the final world space. The total loss function is defined as a weighted combination of multiple losses specific to the task:

$$L_{\text{total}} = \lambda_d L_d + \lambda_t L_t + \lambda_r L_r + \lambda_{\text{shape}} L_{\text{shape}} + \lambda_{3D} L_{3D}, \quad (11)$$

where $L_{\text{total}}$ is the total loss, a weighted sum of the 2D detection loss $L_{\text{d}}$, translation loss $L_{\text{t}}$, rotation loss $L_{\text{r}}$, shape loss $L_{\text{shape}}$, and the 3D spatial loss $L_{\text{3D}}$, with each loss component being scaled by its respective weight $\lambda_d$, $\lambda_t$, $\lambda_r$, $\lambda_{shape}$ and $\lambda_{3D}$.

# Experiments

## Dataset and Metric

The proposed framework is trained and evaluated on the ApolloCar3D dataset, a publicly available dataset that provides each vehicle instance with an industry-grade, real-scale 3D CAD model and dense semantic keypoints. This dataset is particularly suitable for monocular vehicle pose and shape reconstruction as precise 3D shape priors and well-defined geometric constraints are provided through its high-fidelity CAD models, while the dense keypoint annotations enable effective supervision of structural details. Following previous methods (Lee et al. 2023), the dataset is split into 4077 training and 200 validation images.

The evaluation utilizes the instance 3D average precision (A3DP) metrics from (Song et al. 2019), which employs 10 thresholds (criteria from loose to strict) to jointly measure translation, rotation, and 3D car shape reconstruction accuracy. The results on the loose and strict criteria are denoted as c-l and c-s, respectively. During evaluation, Euclidean distance is used for 3D translation, while arccos distance is used for 3D rotation. For 3D shape reconstruction, a predicted mesh is rendered into 100 views to compute the mean IoU (mIoU) with the ground truth masks. In addition to the absolute distance error, the relative error in translation is also evaluated to emphasize model performance on nearby cars, which are more critical for autonomous driving. The A3DP metrics evaluate in their relative and absolute versions are denoted as A3DP-Rel and A3DP-Abs, respectively.

## Implementation Details

The model is constructed based on the Mask R-CNN framework and employs a two-stage training strategy. The first stage is consistent with BAAM (Lee et al. 2023), utilizing a pre-trained model from BAAM to establish a robust 2D perception foundation. The second stage involves end-to-end fine-tuning with a complete multi-task objective function. The weights are set to $\lambda_d = 1.0$, $\lambda_t = 0.5$, $\lambda_r = 1.0$, $\lambda_{\text{shape}} = 3.0$, and $\lambda_{3D} = 0.01$, same as BAAM. Shape loss is given the highest weight to prioritize the generation of high-fidelity geometric structures. The model uses the AdamW optimizer with a global batch size of 4. The learning rate is initialized at $1 \times 10^{-4}$ and is decayed to $1 \times 10^{-5}$ for the final 10 epochs. All experiments are performed on two NVIDIA RTX A6000 GPUs.

## Comparison with State-of-the-Art Approaches

A comprehensive quantitative evaluation of MonoVPR is conducted on the ApolloCar3D dataset, comparing its performance against multiple state-of-the-art methods, including DeepMANTA (Chabot et al. 2017), Keypoints-based (Song et al. 2019), 3D-RCNN (Kundu, Li, and Rehg 2018), Directed-based (Song et al. 2019), GSNet (Ke et al. 2020), and BAAM (Lee et al. 2023).

To ensure a fair comparison, the BAAM baseline is re-implemented and evaluated under identical experimental conditions. As summarized in Table 1, MonoVPR consistently surpasses all compared approaches, including the re-implemented BAAM†. Specifically, the A3DP-Abs mean score is improved from 24.20 to 25.60, while the A3DP-Rel mean increases from 21.93 to 23.26. These results demonstrate the effectiveness and superiority of the proposed framework.

## Ablation Study

**Incremental Integration of Modules** Table 2 shows the incremental gains over the baseline. The BIMR module further improves performance by iteratively refining vertices, boosting A3DP-Abs by 3.6% and A3DP-Rel by 5.7% over the baseline. The combined HDCA+BIMR model achieves the best results, with total gains of 5.8% (A3DP-Abs) and 6.1% (A3DP-Rel), validating the approach for improving pose and shape reconstruction.

| Method | Detailed shape | A3DP-Abs | | | A3DP-Rel | | |
|---|---|---|---|---|---|---|---|
| | | mean | c-l | c-s | mean | c-l | c-s |
| DEEPMETA | ✗ | 20.10 | 30.69 | 23.76 | 16.04 | 23.76 | 19.80 |
| Keypoints-based | ✗ | 20.40 | 31.68 | 24.75 | 16.53 | 24.75 | 19.80 |
| 3D-RCNN | ✓ | 16.44 | 29.70 | 19.80 | 10.79 | 17.82 | 11.88 |
| Directed-based | ✓ | 15.15 | 28.71 | 17.82 | 11.49 | 17.82 | 11.88 |
| GSNet | ✓ | 18.91 | 37.42 | 18.36 | 20.21 | 40.50 | 19.85 |
| BAAM(reported) | ✓ | 25.19 | 47.31 | 23.13 | 22.85 | 46.21 | 20.31 |
| BAAM† | ✓ | 24.20 | 45.82 | 22.67 | 21.93 | 44.97 | 19.27 |
| MonoPVR(Ours) | ✓ | **25.60** | **48.66** | **24.42** | **23.26** | **46.66** | **20.73** |

Table 1: Comparison with the SOTA methods on the ApolloCar3D dataset. (†) means reproduced.

| Method | A3DP-Abs | | | A3DP-Rel | | |
|---|---|---|---|---|---|---|
| | mean | c-l | c-s | mean | c-l | c-s |
| Baseline | 24.20 | 45.82 | 22.67 | 21.93 | 44.97 | 19.27 |
| +HDCA | 24.72 | 46.88 | 23.07 | 22.88 | 46.82 | 19.77 |
| +BIMR | 25.06 | 47.03 | 23.01 | 23.18 | 46.19 | 20.47 |
| +HDCA+BIMR | **25.60** | **48.66** | **24.42** | **23.26** | **46.66** | **20.73** |

Table 2: Results on ApolloCar3D with different combinations of MonoVPR.

| Method | A3DP-Abs | | | A3DP-Rel | | |
|---|---|---|---|---|---|---|
| | mean | c-l | c-s | mean | c-l | c-s |
| Baseline | 24.20 | 45.82 | 22.67 | 21.93 | 44.97 | 19.27 |
| BIMR(t=1) | 25.35 | 47.84 | 23.27 | 23.21 | 46.55 | 19.89 |
| BIMR(t=2) | 25.20 | 47.08 | 23.98 | 23.08 | 46.06 | 20.37 |
| BIMR(t=3) | **25.60** | **48.66** | **24.42** | **23.26** | **46.66** | **20.73** |
| BIMR(t=4) | 25.22 | 46.45 | 24.28 | 23.12 | 46.57 | 20.12 |

Table 3: Results on ApolloCar3D with different iteration times of BIMR components.

**Optimal Iteration Count for BIMR** Table 3 shows the ablation study on the number of iterative loops for BIMR. Performance peaks at three iterations, corresponding to a favorable trade-off for detailed shape recovery. Four iterations lead to performance degradation and begin to compromise the refined geometry.

Figure 5 visualizes the iterative refinement process. The initial step (Iter. 1) corrects global inaccuracies, while subsequent steps (Iter. 2, Iter. 3) focus refinement on intricate surfaces for detailed tuning. This tanh-bounded correction process visually confirms the module's ability to resolve topological artifacts and achieve physically plausible geometry.

**Gated Dual-Path Attention in HDCA** As shown in Table 4, the ablation study validates the necessity of the GDA's dual-branch design. The full model (HDCA+GDA) achieves the best performance, while using only LDP or GGP individually results in performance degradation. This confirms that

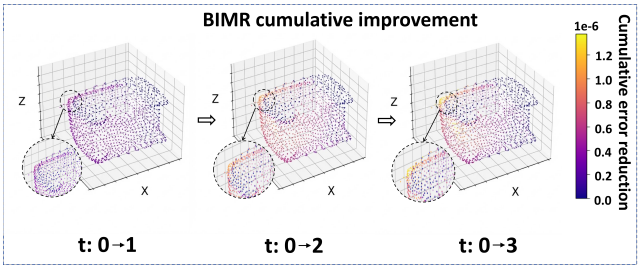


Figure 5: Qualitative analysis of the progressive refinement process in the BIMR module. The three subplots depict the cumulative error reduction of the vehicle mesh after one, two, and three iterations, respectively, with each benchmarked against the initial coarse prediction. The color of each vertex corresponds to the magnitude of its error reduction, where brighter tones progressing towards yellow denote a greater degree of geometric correction.

robust pose estimation requires the synergy of both local fidelity from LDP and global awareness from GGP.

**Performance Comparison at Various Scales** Quantitative analysis in Table 5 validates the effect of HDCA in resolving scale-dependent degradation. MonoVPR surpasses the baseline BAAM across all metrics, with the most pronounced improvement for challenging distant 'S' objects. This is because the HDCA module fuses scene semantics with object cues, compensating for sparse features to enable more precise 3D inference.

| Method | A3DP-Abs | | | A3DP-Rel | | |
|---|---|---|---|---|---|---|
| | mean | c-l | c-s | mean | c-l | c-s |
| HDCA(w/o GDA) | 25.41 | 47.88 | 23.85 | 23.01 | 45.85 | 20.63 |
| HDCA+LDP | 24.19 | 46.50 | 22.27 | 21.92 | 45.22 | 18.64 |
| HDCA+GCP | 25.05 | 47.45 | 23.50 | 22.90 | 46.51 | 19.53 |
| HDCA+GDA | **25.60** | **48.66** | **24.42** | **23.26** | **46.66** | **20.73** |

Table 4: Results on ApolloCar3D with different attentions of HDCA.


Figure 6: Qualitative comparison results of the proposed method with BAAM on the ApolloCar3D dataset.

| Scale | RE (deg) ↓ | | TE (m) ↓ | | mIoU ↑ | |
|---|---|---|---|---|---|---|
| | B | M | B | M | B | M |
| All | 6.24 | **6.01** | 1.28 | **1.27** | 89.01 | **89.43** |
| S | 9.64 | **8.15** | 1.61 | **1.58** | 85.99 | **87.62** |
| M | 7.45 | **6.99** | 1.52 | **1.50** | **89.05** | 88.86 |
| L | 5.45 | **5.38** | 1.14 | **1.13** | 89.71 | **89.82** |

Table 5: The performance comparison between BAAM (B) and MonoVPR (M). 'All' denotes the overall average performance calculated across all distance scales. 'RE' and 'TE' stand for rotation error and translation error. 'S', 'M' and 'L' stand for small(d $>$ 50m), medium(20m $<$ d $<$ 50m) and large(d $<$ 20m). 'd' stands for distances.

| Model | GPU Memory (G) | Params (M) |
|---|---|---|
| BAAM | **10.22** | **130.99** |
| MonoVPR | 11.51 | 131.16 |

Table 6: Computational Comparison.

**Computational Comparison** Computational efficiency is further analyzed in Table 6. The overall framework introduces minimal overhead relative to the baseline, with total model parameters increasing marginally from 130.99M to 131.16M and GPU memory usage rising from 10.22G to 11.51G. The BIMR module contains only 0.93M parameters and consuming 0.78G of GPU memory. This modest computational cost is well justified by a performance improvement on key metrics, demonstrating a favorable trade-off between accuracy and efficiency for robust vehicle reconstruction in autonomous driving applications.

**Analysis of Shape Fusion Weights** Table 7 shows that learnable fusion weights outperform fixed weights on nearly all key metrics. This confirms that dynamically balancing the two shape components ($M_{\text{base}}$ and $O^{(T)}$) is crucial for accurate reconstruction.

| Method | A3DP-Abs | | | A3DP-Rel | | |
|---|---|---|---|---|---|---|
| | mean | c-l | c-s | mean | c-l | c-s |
| Fixed weights | 25.04 | 46.91 | 23.11 | 23.20 | **46.81** | 20.25 |
| Learnable weights | **25.60** | **48.66** | **24.42** | **23.26** | 46.66 | **20.73** |

Table 7: Ablation study on the shape fusion weights in BIMR. The table compares 'Fixed weights' (where $w_1 = w_2 = 1$) against 'Learnable weights' (where $w_1, w_2$ are optimized end-to-end).

## Qualitative Analysis

As depicted in Figure 6, qualitative comparison on ApolloCar3D visually validates MonoVPR achieves accurate pose estimation via HDCA's dynamic fusion of scene and object cues, and eliminates hollowness via BIMR's progressive, bounded correction, producing coherent meshes. This synergy of context adaptation and geometry refinement overcomes prior limitations.

## Conclusion

This paper presents MonoVPR, a framework for monocular 3D vehicle pose and shape reconstruction that mitigates geometric ambiguity and structural hollowness through dynamic context adaptation and iterative mesh refinement. The HDCA module mitigates scale-dependent degradation by integrating multi-scale object geometry and scene semantics. The BIMR module enhances shape recovery by progressively optimizing template deformations via multi-head attention and a bounded correction mechanism, yielding realistic reconstructions. Experiments on ApolloCar3D show SOTA performance, especially for long-range scenarios. However, several limitations remain, including a two-stage pipeline dependent on external 2D detectors. Future work includes developing an end-to-end framework and investigating domain adaptation for illumination robustness.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62001400, 52441801, 61802053, and 62372387), Natural Science Foundation of Sichuan Province (Grant No. 2024NSFSC0494 and 2024NSFSC0508), Fundamental Research Funds for the Central Universities (2682024ZTPY044, 2682025ZD004), China Postdoctoral Science Foundation (Grant No. 2021M702713), Special Research Funding under Yibin Municipal-University Dual Agreement (Grant No. YBSCXY2024010012 and YBSCXY2024010006), and the Fund of National Laboratory on Adaptive Optics, China, (Grant No. FNLAO-24-ZD-O02).

## References

Bai, X.; Hu, Z.; Zhu, X.; Huang, Q.; Chen, Y.; Fu, H.; and Tai, C.-L. 2022. TransFusion: Robust LiDAR-Camera Fusion for 3D Object Detection with Transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1090–1099.

Chabot, F.; Chaouch, M.; Rabarisoa, J.; Teuliere, C.; and Chateau, T. 2017. Deep MANTA: A Coarse-to-fine Many-Task Network for joint 2D and 3D vehicle analysis from monocular image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2040–2049.

Cho, J.; Youwang, K.; Yang, H.; and Oh, T.-H. 2025. Robust 3D Shape Reconstruction in Zero-Shot from a Single Image in the Wild. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 22786–22798.

Erçelik, E.; Yurtsever, E.; Liu, M.; Yang, Z.; Zhang, H.; Topçam, P.; Listl, M.; Caylı, Y. K.; and Knoll, A. 2022. 3D Object Detection with a Self-supervised Lidar Scene Flow Backbone. In *European Conference on Computer Vision*, 247–265.

Gao, S.-H.; Cheng, M.-M.; Zhao, K.; Zhang, X.-Y.; Yang, M.-H.; and Torr, P. 2019. Res2Net: A New Multi-scale Backbone Architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2): 652–662.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, 2961–2969.

He, T.; and Soatto, S. 2019. Mono3D++: Monocular 3D Vehicle Detection with Two-Scale 3D Hypotheses and Task Priors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8409–8416.

Huang, Z.; Stojanov, S.; Thai, A.; Jampani, V.; and Rehg, J. M. 2024. ZeroShape: Regression-based Zero-shot Shape Reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10061–10071.

Ke, L.; Li, S.; Sun, Y.; Tai, Y.-W.; and Tang, C.-K. 2020. GSNet: Joint Vehicle Pose and Shape Reconstruction with Geometrical and Scene-aware Supervision. In *European Conference on Computer Vision*, 515–532.

Kundu, A.; Li, Y.; and Rehg, J. M. 2018. 3D-RCNN: Instance-level 3D Object Reconstruction via Render-and-Compare. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3559–3568.

Lee, H.-J.; Kim, H.; Choi, S.-M.; Jeong, S.-G.; and Koh, Y. J. 2023. BAAM: Monocular 3D Pose and Shape Reconstruction with Bi-Contextual Attention Module and Attention-Guided Modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9011–9020.

Li, P.; Chen, X.; and Shen, S. 2019. Stereo R-CNN based 3D object detection for Autonomous Driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7644–7652.

Li, W.; Liu, S.; Qiao, P.; and Dou, Y. 2025. Mono3R: Exploiting Monocular Cues for Geometric 3D Reconstruction. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 11081–11090.

Li, Z.; Xu, X.; Lim, S.; and Zhao, H. 2024. UniMODE: Unified Monocular 3D Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 16561–16570.

Liu, Z.; Zhou, D.; Lu, F.; Fang, J.; and Zhang, L. 2021. AUTOSHAPE: Real-Time Shape-Aware Monocular 3D Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 15641–15650.

Lu, Y.; Ma, X.; Yang, L.; Zhang, T.; Liu, Y.; Chu, Q.; Yan, J.; and Ouyang, W. 2021. Geometry Uncertainty Projection Network for Monocular 3D Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 3111–3121.

Manhardt, F.; Kehl, W.; and Gaidon, A. 2019. ROI-10D: Monocular Lifting of 2D Detection to 6D Pose and Metric Shape. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2069–2078.

Murhij, Y.; and Yudin, D. 2024. DAGM-MONO: Deformable Attention-guided Modeling for Monocular 3D Reconstruction. *Optical Memory and Neural Networks*, 33(2): 144–156.

Park, D.; Ambrus, R.; Guizilini, V.; Li, J.; and Gaidon, A. 2021. Is Pseudo-Lidar needed for Monocular 3D Object Detection? In *Proceedings of the IEEE International Conference on Computer Vision*, 3142–3152.

Peng, L.; Xu, J.; Cheng, H.; Yang, Z.; Wu, X.; Qian, W.; Wang, W.; Wu, B.; and Cai, D. 2024. Learning Occupancy for Monocular 3D Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10281–10292.

Pu, F.; Wang, Y.; Deng, J.; and Yang, W. 2025. MonoDGP: Monocular 3D Object Detection with Decoupled-Query and Geometry-Error Priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 6520–6530.

Ranasinghe, Y.; Hegde, D.; and Patel, V. M. 2024. MonoDiff: Monocular 3D Object Detection and Pose Estimation with Diffusion Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10659–10670.

Shi, P.; Dong, X.; Ge, R.; Liu, Z.; and Yang, A. 2025. DP-M3D: Monocular 3D Object Detection Algorithm with Depth Perception Capability. *Knowledge-Based Systems*, 318: 113539.

Shi, S.; Guo, C.; Jiang, L.; Wang, Z.; Shi, J.; Wang, X.; and Li, H. 2020. PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10529–10538.

Song, X.; Wang, P.; Zhou, D.; Zhu, R.; Guan, C.; Dai, Y.; Su, H.; Li, H.; and Yang, R. 2019. ApolloCar3D: A Large 3D Car Instance Understanding Benchmark for Autonomous Driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5452–5462.

Sun, J.; Chen, L.; Xie, Y.; Zhang, S.; Jiang, Q.; Zhou, X.; and Bao, H. 2020. Disp R-CNN: Stereo 3D Object Detection via Shape Prior Guided Instance Disparity Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10548–10557.

Tan, M.; Pang, R.; and Le, Q. V. 2020. EFFICIENTDET: Scalable and Efficient Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10781–10790.

Wang, Y.; Chao, W.-L.; Garg, D.; Hariharan, B.; Campbell, M.; and Weinberger, K. Q. 2019. Pseudo-LiDAR from Visual Depth Estimation: Bridging the Gap in 3D Object Detection for Autonomous Driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8445–8453.

Wu, Z.; Wu, Y.; Pu, J.; Li, X.; and Wang, X. 2023. Attention-Based Depth Distillation with 3D-Aware Positional Encoding for Monocular 3D Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2892–2900.

Yan, L.; Yan, P.; Xiong, S.; Xiang, X.; and Tan, Y. 2024. MonoCD: Monocular 3D Object Detection with Complementary Depths. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10248–10257.

Zhang, H.; Tian, Y.; Zhou, X.; Ouyang, W.; Liu, Y.; Wang, L.; and Sun, Z. 2021. PyMAF: 3D Human Pose and Shape Regression with Pyramidal Mesh Alignment Feedback Loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11446–11456.

Zhang, Y.; Lu, J.; and Zhou, J. 2021. Objects are Different: Flexible Monocular 3D Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3289–3298.

Zhao, T.; Chen, Y.; Wu, Y.; Liu, T.; Du, B.; Xiao, P.; Qiu, S.; Yang, H.; Li, G.; Yang, Y.; et al. 2024. Improving Bird's Eye View Semantic Segmentation by Task Decomposition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 15512–15521.