PROTODN: APPEARANCE-AWARE DENOISING WITH DYNAMIC VISUAL PROTOTYPES FOR MONOCULAR 3D OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Monocular 3D object detection is inherently challenging due to the depth ambiguity and incomplete observations from a single view. Recent denoising-based monocular 3D detectors focus on improving query learning and training stability. However, their denoising strategies remain largely centered on distribution modeling and geometric guidance, leaving instance-level visual information underexploited and limiting the modeling of challenging objects under occlusion, long-range observation, and truncation. To address this issue, we propose ProtoDN, an appearance-aware denoising framework with dynamic visual prototypes. We first introduce a dynamic visual prototype memory to aggregate class-specific visual priors. Building on this memory, we develop two core modules: (i) Geometry–Visual Joint Difficulty Modeling integrates geometric and visual cues to estimate instance difficulty and modulates denoising perturbations accordingly, enabling visually informed difficulty-aware denoising. (ii) Prototype-Guided Visual Enhancement leverages visual prototypes as instance-relevant appearance priors and explicitly injects them into denoising representations, compensating for insufficient visual information during reconstruction. Extensive experiments demonstrate superior performance of our method over recent approaches, with particularly notable improvements on challenging instances.

1 INTRODUCTION

Monocular 3D object detection aims to recover the 3D location, dimensions, and orientation of objects from a single RGB image. Compared with LiDAR-based (Shi et al., 2023; Wu et al., 2023a; Peng et al., 2024; Liu et al., 2025) or stereo-based (Wu et al., 2023b; Chen et al., 2022; Li et al., 2024; Chen et al., 2024) perception systems, monocular approaches offer an attractive alternative due to their simple sensor configuration and low deployment cost. However, monocular 3D detection inherently suffers from ill-posed depth estimation (Liu et al., 2020; Lu et al., 2021; Kumar et al., 2025) because a single image provides neither direct depth measurements nor multi-view geometric constraints. The lack of viewpoint disparity limits reliable geometric reasoning for accurate 3D localization, while occluded, truncated, or partially observed objects further increase the ambiguity of monocular perception.

To alleviate the inherent challenges of monocular perception, an increasing number of monocular 3D detectors have adopted Transformer-based architectures, particularly the DEtection TRansformer (DETR) (Carion et al., 2020), to strengthen global reasoning and spatial understanding. Representative methods such as MonoDETR (Zhang et al., 2023b) and MonoDGP (Pu et al., 2025) further enhance query-based reasoning and geometric modeling, while subsequent studies improve model robustness through masked representation learning (Jiang et al., 2024), foreground-aware supervision (Wu et al., 2024), geometric consistency (Wang et al., 2026), and inter-attribute dependency modeling (Zhang et al., 2026). However, these DETR-based monocular 3D detection frameworks still suffer from training issues such as unstable bipartite matching and difficult query optimization.

To address these training issues, several denoising-based methods (Li et al., 2022a; Zhang et al., 2023a) have shown that introducing query denoising during training can provide more direct supervision for query learning and improve training stability in 2D object detection. Inspired by the effec-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

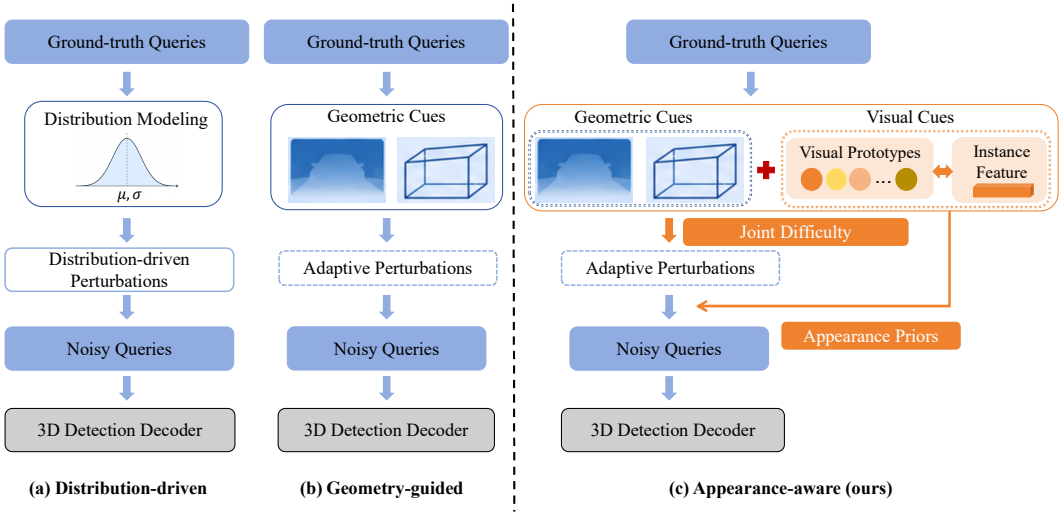


Figure 1: Comparison of denoising paradigms for monocular 3D object detection: (a) Distribution-driven denoising with distribution-driven perturbations, (b) Geometry-guided denoising with geometry-adaptive perturbations, and (c) Appearance-aware denoising (ours) with joint difficulty-aware perturbation and appearance priors injection.

tiveness of query denoising, recent studies have extended denoising mechanisms to monocular 3D object detection with dedicated noise construction strategies. These approaches can be broadly categorized into distribution-driven and geometry-guided denoising. As shown in Fig. 1(a), distribution-driven methods generate noisy queries through distribution modeling to facilitate denoising learning and query optimization (Vu et al., 2025; 2026). In contrast, geometry-guided methods in Fig. 1(b) adaptively modulate query perturbations using geometric cues, providing more targeted geometric supervision during denoising (Lee et al., 2026). Despite their different noise construction and modulation strategies, existing methods mainly focus on distribution modeling and geometric guidance, while instance-level visual information remains underexplored, limiting the modeling of challenging instances under occlusion and truncation.

To address this gap, we propose ProtoDN, an appearance-aware denoising framework with dynamic visual prototypes. We first construct a dynamic visual prototype memory to maintain visual priors that remain aligned with the evolving representation space. Building on this memory, we incorporate instance-level visual information into denoising learning through two complementary mechanisms, as illustrated in Fig. 1(c). **Geometry–Visual Joint Difficulty Modeling** jointly exploits geometric and visual cues to estimate instance-level detection difficulty and adaptively modulates denoising perturbations accordingly, thereby incorporating visual information implicitly into denoising learning. **Prototype-Guided Visual Enhancement** retrieves instance-relevant appearance priors from the visual prototype memory and directly injects them into denoising representations, enabling explicit integration of visual information.

Our contributions are summarized as follows:

- Geometry–Visual Joint Difficulty Modeling is introduced to jointly exploit geometric and visual cues for instance-level difficulty estimation, addressing the limitation of geometry-dominated difficulty modeling.
- Prototype-Guided Visual Enhancement is proposed to leverage dynamic visual prototypes as instance-relevant appearance priors, alleviating the lack of instance-level visual information during denoising.
- Extensive experiments on the KITTI dataset demonstrate the effectiveness of ProtoDN, yielding significant performance improvements with particularly notable gains on challenging instances.

2 Related Work

Monocular 3D Object Detection. Monocular 3D object detection aims to recover 3D object properties from a single RGB image, but remains inherently challenging due to the absence of direct depth measurements. Early approaches exploit geometric constraints between the 2D image plane and 3D space (Mousavian et al., 2017; Naiden et al., 2019), while subsequent methods improve monocular spatial reasoning through uncertainty-aware depth estimation, geometric priors, and complementary depth cues (Lu et al., 2021; Zhang et al., 2021; Li et al., 2022b; Yan et al., 2024). More recently, Transformer-based methods have further advanced query-based reasoning and global relation modeling. MonoDETR (Zhang et al., 2023b) introduces a depth-guided Transformer that leverages scene-level depth information to enhance interactions between object queries and image features. MonoDGP (Pu et al., 2025) further improves 3D attribute modeling using decoupled queries and geometry-error priors. Subsequent works further explore more fine-grained modeling strategies. SPAN (Wang et al., 2026) enhances geometric consistency through spatial and projection alignment. MonoCoP (Zhang et al., 2026) models inter-attribute correlations and coordinates chain-based and parallel prediction with uncertainty guidance. MonoCLUE (Yang et al., 2026) strengthens local object representations through object-aware clustering. Although these methods have achieved notable progress, they still insufficiently account for variations across individual object instances, especially under occlusion, long-range observation, and truncation. Our method alleviates this limitation by incorporating instance-level visual information to better model diverse object instances.

Denoising-based Object Detection. Denoising training has become an effective strategy for stabilizing the optimization of DETR-style detectors. DN-DETR (Li et al., 2022a) introduces noisy ground-truth queries and trains the decoder to reconstruct their clean targets, alleviating unstable bipartite matching and accelerating convergence. DINO (Zhang et al., 2023a) further extends this paradigm with contrastive denoising to improve query learning. Denoising strategies have also been introduced into 3D object detection, where perturbed ground-truth queries provide additional spatial supervision during training (Liu et al., 2024). In monocular 3D object detection, MonoVQD (Vu et al., 2025) introduces variational query denoising to mitigate gradient vanishing by injecting stochasticity into the denoising process. Mono3DV (Vu et al., 2026) combines 3D-aware bipartite matching with variational query denoising to stabilize training when incorporating 3D attributes. MonoDLGD (Lee et al., 2026) adaptively adjusts label perturbation magnitudes using geometry-related and depth-related information, providing more targeted geometric supervision for denoising training. Despite their effectiveness, existing denoising methods remain largely centered on distribution modeling and geometric information, leaving instance-level appearance cues underexplored. Our method addresses this limitation through joint difficulty-aware perturbation and appearance priors injection, enabling more effective exploitation of visual information.

Prototype-based Representation Learning. Prototype-based learning represents recurring visual patterns with compact feature references and has been widely used to improve representation consistency and discrimination (Hartigan & Wong, 1979; Caron et al., 2018; Snell et al., 2017; Caron et al., 2020). Recent monocular 3D detection methods have also explored prototype-like visual priors. MonoCLUE (Yang et al., 2026) employs object-level clustering and scene memory to enhance appearance representations, while MonoSAOD (Jung et al., 2026) combines prototype similarity with depth uncertainty for reliable pseudo-label selection. In contrast to using prototypes primarily for representation enhancement or pseudo-label selection, we employ a dynamic visual prototype memory within the denoising pathway, where it serves as a unified visual reference for instance-adaptive perturbation control and query enhancement.

3 Method

3.1 Overview

As shown in Fig. 2(a), we follow the denoising-based monocular 3D detection paradigm and focus on constructing noisy queries through two complementary components: Geometry–Visual Joint Difficulty Modeling in Fig. 2(b) and Prototype-Guided Visual Enhancement in Fig. 2(c). Fig. 2(d) illustrates the construction of the Dynamic Visual Prototype Memory, which provides visual priors

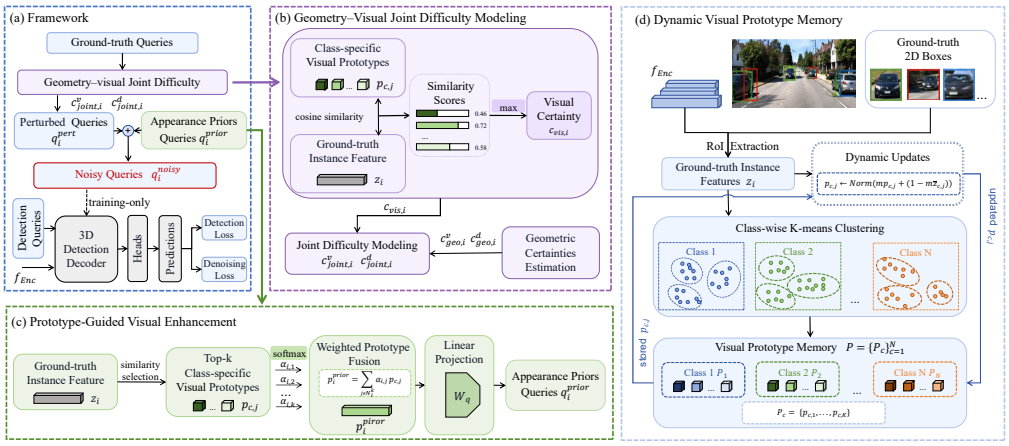


Figure 2: The overall framework of ProtoDN. (a) Framework: denoising-based monocular 3D detection with training-only noisy queries. (b) Geometry–Visual Joint Difficulty Modeling: geometric and visual certainty are fused into a joint difficulty estimate. (c) Prototype-Guided Visual Enhancement: class-specific visual prototypes are aggregated via Top-k Soft Fusion to construct appearance priors. (d) Dynamic Visual Prototype Memory: the visual prototype memory is constructed and refined through dynamic updates.

for both components. All proposed components are used only during training, while the inference pipeline remains unchanged.

3.2 Dynamic Visual Prototype Memory

To obtain rich and representative visual priors, we introduce a Dynamic Visual Prototype Memory to effectively aggregate class-specific visual patterns. Specifically, given the encoded feature f_{Enc}, we extract the region-of-interest (RoI) feature of each object instance according to its ground-truth 2D bounding box $\mathbf{b}_i$. The normalized instance feature $\mathbf{z}_i$ is obtained as:

$$\mathbf{z}_i = \text{Norm}\left(\phi_{\text{roi}}\left(f_{\text{Enc}}, \mathbf{b}_i\right)\right) \in \mathbb{R}^C, \tag{1}$$

where $\phi_{\text{roi}}(\cdot)$ denotes the RoI feature extraction process, including feature cropping, adaptive average pooling, spatial aggregation, and feature projection, and C denotes the visual prototype feature dimension.

The resulting instance features are grouped according to their class labels y_i. For each class c, the corresponding instance features are collected as

$$\mathcal{Z}_c = \{\mathbf{z}_i \mid y_i = c\}, \tag{2}$$

We then perform K-means clustering on $\mathcal{Z}_c$ to construct the class-specific prototype memory:

$$\mathbf{P}_c = \text{KMeans}\left(\mathcal{Z}_c, K\right) = \{\mathbf{p}_{c,j}\}_{j=1}^{K}, \tag{3}$$

where K denotes the number of visual prototypes per class. Each prototype $\mathbf{p}_{c,j}$ corresponds to the center of the j-th cluster and is initialized as the mean feature of the samples assigned to that cluster. Finally, the class-specific prototype memories are combined to form the complete visual prototype memory $\mathbf{P} = \{\mathbf{P}_c\}_{c=1}^{N}$, where N denotes the total number of classes.

To adapt to the evolving feature space, we update the visual prototype memory online using the ground-truth instance features extracted from the current mini-batch. For the i-th instance belonging to class c, we compute its update weight according to the similarity $s_{i,j}$ with prototype $\mathbf{p}_{c,j}$:

$$s_{i,j} = \mathbf{z}_i^\top \mathbf{p}_{c,j}, \qquad a_{i,j} = \frac{\exp(s_{i,j})}{\sum_{m=1}^{K} \exp(s_{i,m})}, \qquad j = 1, \ldots, K. \tag{4}$$

For prototype j of class c, the aggregated current-batch feature $\bar{\mathbf{z}}_{c,j}$ is computed as:

$$\bar{\mathbf{z}}_{c,j} = \frac{\sum_{i:y_i=c} a_{i,j}\mathbf{z}_i}{\sum_{i:y_i=c} a_{i,j}}, \qquad \text{if } \sum_{i:y_i=c} a_{i,j} > \epsilon, \tag{5}$$

where ϵ is a small constant used to avoid unstable updates with insufficient assignments. Each valid prototype is updated using exponential moving average:

$$\mathbf{p}_{c,j} \leftarrow \mathrm{Norm}\left(m\mathbf{p}_{c,j} + (1-m)\bar{\mathbf{z}}_{c,j}\right), \tag{6}$$

where $m \in [0, 1)$ is the prototype momentum. Through the above dynamic updates, the visual prototypes continuously adapt to the evolving feature space, maintaining representative class-specific visual information. The resulting prototype memory serves as a shared visual prior for the modules in the subsequent sections.

3.3 Geometry–Visual Joint Difficulty Modeling

To more comprehensively characterize the detection variations across different instances, we propose to model instance-level detection difficulty using geometric and visual cues. First, we estimate the visual certainty $\mathbf{c}_{vis,i}$ by measuring the matching degree between the instance feature $\mathbf{z}_i$ and the class-specific visual prototypes $\mathbf{p}_{c,j}$:

$$c_{\mathrm{vis},i} = \mathrm{CLIP}_{(0,1)}\left(\frac{1 + \max_{j\in\{1,\ldots,K\}} \mathbf{z}_i^\top \mathbf{p}_{c,j}}{2}\right). \tag{7}$$

where $\mathrm{CLIP}_{(0,1)}(x)$ clamps the input into the interval $[0, 1]$. Subsequently, we fuse the visual certainty with the geometric certainty defined by MonoDLGD (Lee et al., 2026) to obtain the joint certainties:

$$c^a_{\mathrm{joint},i} = c^a_{\mathrm{geo},i}\left[(1-\lambda_t) + \lambda_t c_{\mathrm{vis},i}\right], \qquad a \in \{v, d\}. \tag{8}$$

where v and d denote the projected-box attribute and depth, respectively. The geometric certainties $c^v_{\mathrm{geo},i}$ and $c^d_{\mathrm{geo},i}$ are derived from the predicted log-variances of the projected box and depth through an inverse exponential mapping. λ_t denotes the visual ramp-up coefficient that controls the contribution of visual certainty during training.

$$\lambda_t = \begin{cases} 0, & t < E_w, \\ \min\left(1, \frac{t-E_w}{E_r}\right), & t \geq E_w, \end{cases} \tag{9}$$

where t denotes the current training epoch, E_w is the warm-up epoch number, and E_r controls the duration of the linear ramp-up stage.

These joint certainties characterize instance-level detection difficulty and are then used to adaptively modulate the perturbation magnitudes of the corresponding attributes:

$$\begin{aligned} \tilde{x}^v_i &= \mathrm{CLIP}_{(0,1)}\left(x^v_i + o^v_i c^v_{\mathrm{joint},i}\delta^v_i\right), \quad v \in \{l, r\}, \\ \tilde{y}^v_i &= \mathrm{CLIP}_{(0,1)}\left(y^v_i + o^v_i c^v_{\mathrm{joint},i}\delta^v_i\right), \quad v \in \{t, b\}, \\ \tilde{d}_i &= d_i + d_i c^d_{\mathrm{joint},i}\delta^d_i. \end{aligned} \tag{10}$$

where x^v_i and y^v_i denote the horizontal and vertical projected box coordinates, respectively. l, r, t, and b correspond to the left, right, top, and bottom sides of the 2D box. d_i denotes the ground-truth depth, and o^v_i is the corresponding box offset used to scale the spatial perturbation. δ^v_i and δ^d_i are perturbation terms that absorb the effects of sampled noise and perturbation strength.

Then, the perturbed ground-truth attributes in Eq. 10 are encoded by a three-layer MLP to form the perturbed queries $\mathbf{q}^{\mathrm{pert}}_i$. This allows the perturbation strength to better match the actual detection difficulty of each instance, alleviating the limitation of geometry-only difficulty modulation while implicitly incorporating instance-level visual information into denoising learning. These perturbed queries are then passed to the module introduced in Sec. 3.4, which further incorporates appearance information before 3D detection.

3.4 PROTOTYPE-GUIDED VISUAL ENHANCEMENT

To further introduce explicit visual information into the denoising process, we propose to aggregate class-specific visual prototypes via Top-k Soft Fusion to construct appearance priors and inject them into the denoising representations.

For each ground-truth instance i, we first select the top-k most relevant prototypes according to the similarity $s_{i,j}$ between its instance feature $\mathbf{z}_i$ and the same-class prototypes $\mathbf{p}_{c,j}$:

$$\mathcal{N}_i^k = \mathrm{TopK}_j\left(s_{i,j}\right), \tag{11}$$

where $\mathcal{N}_i^k$ denotes the index set of the top-k prototypes for instance i.

We then normalize the similarities within $\mathcal{N}_i^k$ using softmax and aggregate the selected prototypes to obtain the appearance priors:

$$\alpha_{i,j} = \frac{\exp(s_{i,j})}{\sum_{m\in\mathcal{N}_i^k}\exp(s_{i,m})}, \qquad \mathbf{p}_i^{\mathrm{prior}} = \sum_j \alpha_{i,j}\mathbf{p}_{c,j}, \qquad j \in \mathcal{N}_i^k. \tag{12}$$

The aggregated appearance priors are then projected into the query space using a learnable linear transformation:

$$\mathbf{q}_i^{\mathrm{prior}} = \alpha_q \lambda_t c_{\mathrm{vis},i}\mathbf{W}_q\mathbf{p}_i^{\mathrm{prior}}, \tag{13}$$

where α_q controls the overall injection strength. The projected appearance priors are then residually injected into the perturbed queries to obtain the final appearance-aware noisy queries $\mathbf{q}_i^{\mathrm{noisy}}$:

$$\mathbf{q}_i^{\mathrm{noisy}} = \mathbf{q}_i^{\mathrm{pert}} + \mathbf{q}_i^{\mathrm{prior}}. \tag{14}$$

Through residual injection, the resulting $\mathbf{q}_i^{\mathrm{noisy}}$ preserves the original perturbation information while incorporating instance-relevant appearance priors. During training, the appearance-aware noisy queries in Eq. 14 are refined by the 3D detection decoder, and then decoded into 2D boxes and 3D object attributes by the prediction heads. The predictions are then directly supervised by their corresponding ground-truth targets through denoising losses. Finally, the denoising losses are jointly optimized with the standard detection losses to update the model parameters, allowing the instance-level visual information carried by the appearance-aware noisy queries to effectively contribute to model training and improve 3D detection performance.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Dataset. We evaluate our method on the widely used KITTI benchmark (Geiger et al., 2012). KITTI contains 7,481 training images and 7,518 test images covering three object categories: Car, Pedestrian, and Cyclist. Each category is further divided into three difficulty levels, Easy, Moderate, and Hard. Following previous work (Chen et al., 2015), we split the official training set into 3,712 images for training and 3,769 images for validation.

Evaluation Metrics. The detection results are evaluated under three difficulty levels: Easy, Moderate, and Hard. Consistent with prior works, we report average precision in 3D space and bird's-eye view, denoted as AP_{3D} and AP_{BEV}, respectively, which are computed over 40 recall positions according to the established evaluation protocol (Simonelli et al., 2019). The KITTI benchmark ranks all methods mainly based on the Moderate AP_{3D} metric for the Car category.

Implementation Details. We follow the overall framework of MonoDGP (Pu et al., 2025) and MonoDLGD (Lee et al., 2026). ResNet-50 is employed as the backbone to extract multi-scale image features, followed by 2D Transformer processing and depth-aware 3D decoding for 3D prediction, and the training loss settings are consistent with MonoDLGD. All experiments are conducted on a single NVIDIA RTX 4090 GPU with a batch size of 8. For the dynamic visual prototype memory, we set the number of prototypes per class to $K = 5$ and the prototype momentum m for dynamic

Table 1: Comparison with current state-of-the-art methods for the Car category on the KITTI validation and test sets. The best results are highlighted in **bold**, the second-best results are underlined.

Methods	Extra data	Reference	Test, $AP_{BEV\|R40}$			Test, $AP_{3D\|R40}$			Val, $AP_{BEV\|R40}$			Val, $AP_{3D\|R40}$		
			Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
MonoDTR (Huang et al., 2022)	LiDAR	CVPR 2022	28.59	20.38	17.14	21.99	15.39	12.73	33.33	25.35	21.68	24.52	18.57	15.51
DID-M3D (Peng et al., 2022)		ECCV 2022	32.95	22.76	19.83	24.40	16.29	13.75	31.10	22.76	19.50	22.98	16.12	14.03
OccupancyM3D (Peng et al., 2024)		CVPR 2024	35.38	24.18	21.37	25.55	17.02	14.79	35.72	26.60	23.68	26.87	19.96	17.15
MonoPGC (Wu et al., 2023c)	Depth	ICRA 2023	32.50	23.14	20.30	24.68	17.17	14.14	34.06	24.26	20.78	25.67	18.63	15.65
OPA-3D (Su et al., 2023)		RA-L 2023	33.54	22.53	19.22	24.60	17.05	14.25	33.80	25.51	22.13	24.97	19.40	16.59
MonoDETR (Zhang et al., 2023b)	–	ICCV 2023	33.60	22.11	18.60	25.00	16.47	13.58	37.86	26.95	22.80	28.84	20.61	16.38
MonoMAE (Jiang et al., 2024)		NeurIPS 2024	34.15	24.93	21.76	25.60	18.84	16.78	40.26	27.08	23.14	30.29	20.90	17.61
MonoDGP (Pu et al., 2025)		CVPR 2025	35.24	25.23	22.02	26.35	18.72	15.97	39.40	28.20	24.42	30.76	22.34	19.02
SPAN (Wang et al., 2026)		CVPR 2026	35.58	24.83	21.79	27.02	19.30	16.49	40.06	29.17	25.56	30.98	23.26	20.17
MonoCoP (Zhang et al., 2026)		CVPR 2026	36.77	25.57	22.62	27.54	19.11	16.33	42.20	31.29	27.58	32.06	23.98	20.64
MonoCLUE (Yang et al., 2026)		AAAI 2026	36.15	26.15	22.81	27.94	19.70	16.69	41.79	29.91	26.00	**33.74**	24.10	20.58
MonoDLGD (Lee et al., 2026)		AAAI 2026	36.63	25.30	23.13	29.11	19.87	17.74	41.50	30.01	27.43	33.72	24.28	21.05
ProtoDN (Ours)		–	**36.92**	**26.45**	**23.38**	**29.16**	**20.13**	**18.06**	**42.52**	**31.62**	**28.12**	33.56	**24.57**	**22.09**

updating to 0.99. The warm-up epoch number E_w for introducing visual information is set to 40, and the ramp-up duration E_r is set to 10. When constructing the appearance priors, we select the top-k most similar visual prototypes with $k = 3$ and set the query injection strength α_q to 0.1. During training, we adopt the AdamW optimizer (Loshchilov & Hutter, 2019) with an initial learning rate of 2×10^{-4} and train the model for 250 epochs. The learning rate is halved at epochs 85, 125, 165, and 205.

4.2 Main Results

KITTI Validation Results. As shown in Table 1, the proposed method achieves the best performance on most metrics on the KITTI validation set without using any extra data. Compared with the previous best results, it improves $AP_{BEV|R40}$ and $AP_{3D|R40}$ on the Moderate subset by 0.33% and 0.29%, respectively. In particular, the 1.04% improvement in $AP_{3D|R40}$ on the Hard subset demonstrates that our appearance-aware denoising strategy is particularly effective for challenging instances. Notably, our method also outperforms approaches trained with extra data across all metrics. For example, on the Moderate subset, it surpasses OccupancyM3D (Peng et al., 2024) by 4.61% in $AP_{3D|R40}$.

KITTI Leaderboard (Test) Results. As shown in Table 1, ProtoDN consistently achieves state-of-the-art performance across all metrics for both 3D and BEV detection on the KITTI leaderboard. Without using any extra data, ProtoDN improves the previous best Moderate $AP_{BEV|R40}$ and $AP_{3D|R40}$ by 0.30% and 0.26%, respectively. Even compared with approaches trained using additional LiDAR or depth data, ProtoDN maintains superior performance across all test metrics, further demonstrating the effectiveness of our approach.

Efficiency and Performance Comparison Table 2 summarizes the efficiency comparison. Compared with MonoDETR (Zhang et al., 2023b), which has the lowest computational cost among all the compared methods, ProtoDN introduces only 0.32 GFLOPs and 2.63M additional parameters, while improving $AP_{3D|R40}$ and $AP_{BEV|R40}$ on the Moderate subset by 3.96% and 4.67%, respectively. These results demonstrate that ProtoDN achieves a favorable trade-off between detection performance and computational efficiency.

Table 2: Efficiency and performance comparison on the KITTI validation set.

Method	$AP_{3D\|R40}$ ↑	$AP_{BEV\|R40}$ ↑	FLOPs ↓	Params(M) ↓
MonoDETR	20.61	26.95	68.99	37.68
MonoDGP	22.34	28.20	68.99	44.17
MonoCoP	23.98	31.29	71.77	42.50
MonoCLUE	24.10	29.91	72.71	44.17
ProtoDN	**24.57**	**31.62**	69.31	40.31

Qualitative Results. For a more intuitive comparison with the baseline method, we present the 3D and BEV detection results in Fig. 3. It can be observed that the proposed method achieves more



(a) MonoDGP (b) MonoDLGD (c) ProtoDN

Figure 3: Qualitative results on the KITTI validation set. For each example, the left row presents the camera-view visualization, while the right row offers the corresponding bird's-eye view. Ground-truth bounding boxes are shown in green, and predicted boxes are shown in red. Representative objects are highlighted to illustrate the differences between competing methods and our approach.

accurate localization for most objects. The improvement is particularly evident for challenging instances, such as occluded, distant, and truncated objects, further demonstrating that incorporating visual information benefits the localization capability of monocular 3D detectors.

4.3 Ablation Studies

We conduct a series of ablation experiments on the KITTI validation set to evaluate the effectiveness of the proposed components. Following standard practice, $AP_{3D|R40}$ on the Moderate subset is used as the primary evaluation metric.

Component ablation. Table 3 quantitatively evaluates the contribution of each proposed component. The baseline constructs noisy queries using only geometry-based perturbations. With all other settings kept unchanged, introducing Geometry–Visual Joint Difficulty Modeling brings a clear improvement on the Moderate subset, increasing $AP_{BEV|R40}$ by 1.49%, while $AP_{3D|R40}$ also shows a positive gain, confirming the benefit of incorporating visual cues into difficulty-aware perturbation. When dynamic updates are further applied to the visual prototype memory, performance is further improved, especially on challenging instances, where Hard $AP_{3D|R40}$ increases by 0.75%. Based on the dynamically adapted visual prototype memory, Prototype-Guided Visual Enhancement yields further gains on both Moderate and Hard subsets, including a 0.37% improvement in Hard $AP_{3D|R40}$ and a 0.46% gain in Hard $AP_{BEV|R40}$. With all components combined, the full model achieves consistent improvements, demonstrating the complementary effects of the proposed modules.

Table 3: Ablation study on different components of our proposed method. Joint Diff. denotes Geometry–Visual Joint Difficulty Modeling, Visual Enh. denotes Prototype-Guided Visual Enhancement, and Dynamic Upd. denotes the dynamic updates of the visual prototype memory.

Joint Diff.	Visual Enh.	Dynamic Upd.	$AP_{BEV\|R40}$			$AP_{3D\|R40}$		
			Easy	Mod.	Hard	Easy	Mod.	Hard
			41.50	30.01	27.43	**33.72**	24.28	21.05
✓			41.70	31.50	27.72	33.02	24.45	20.97
✓	✓		42.37	31.38	27.95	32.39	24.02	21.02
✓		✓	**42.62**	31.56	27.66	33.65	24.42	21.72
✓	✓	✓	42.52	**31.62**	**28.12**	33.56	**24.57**	**22.09**

Geometry–visual complementarity. To investigate whether visual information provides complementary cues beyond geometry for modeling instance-level detection difficulty, we further analyze the complementarity between geometric and visual certainty. Specifically, we first compute the ge-

Table 4: Geometry–visual complementarity analysis. Match denotes the percentage of ground-truth instances successfully matched with valid predictions, while Depth Err., Center Err., and 3D IoU measure the depth error, 3D center localization error, and 3D overlap quality, respectively.

Group	Match (%)	Depth Err. (m)↓	Center Err. (m)↓	3D IoU↑
High Geo + High Vis	98.85	**0.252**	**0.289**	**0.587**
High Geo + Low Vis	99.19	0.419	0.441	0.469
Low Geo + High Vis	84.19	0.993	1.024	0.262
Low Geo + Low Vis	83.60	1.055	1.098	0.207

ometric and visual certainty of validation samples and divide them into four groups according to the median of each certainty measure. We then report the matching rate and geometric detection metrics, including depth error, 3D center error, and 3D IoU, for each group, as shown in Table 4. Notably, among the two high-geometry groups with similar matching rates, samples with different visual certainty still exhibit clear performance differences. Compared with high-visual-certainty samples, the depth error and 3D center error of low-visual-certainty samples increase from 0.252/0.289 m to 0.419/0.441 m, while the mean 3D IoU decreases from 0.587 to 0.469. These results indicate that geometric cues alone are insufficient to fully characterize instance-level detection difficulty, whereas visual information provides complementary evidence for more accurate difficulty estimation. Therefore, jointly modeling geometric and visual cues provides more reliable instance-level difficulty information for the denoising process and further improves monocular 3D object detection.

Appearance Prior Construction Strategies. To evaluate the effectiveness of Top-k Soft Fusion for appearance prior construction, we compare it with several alternative strategies in Table 5. Cross-Attention takes the perturbed queries as Q and the visual prototypes as K and V to aggregate prototype information. However, since the retrieval is conditioned on already perturbed queries, the resulting prior can be affected by denoising noise, leading to relatively weaker overall performance. Nearest Prototype instead retrieves the single most similar prototype according to the ground-truth instance feature, which provides a more stable instance-specific appearance prior and improves Moderate/Hard $AP_{3D|R40}$ to 24.61/22.12. Top-k Soft Fusion further selects multiple relevant prototypes and combines them using similarity-based softmax weights, allowing complementary appearance patterns to be aggregated rather than relying on a single prototype. This strategy achieves the best $AP_{BEV|R40}$ across all difficulty levels and the highest Easy $AP_{3D|R40}$, while remaining only marginally lower than Nearest Prototype on Moderate and Hard $AP_{3D|R40}$. These results indicate that softly aggregating multiple relevant prototypes provides a richer and more robust appearance prior, and we therefore adopt Top-k Soft Fusion in the final model.

Table 5: Ablation study on different appearance prior construction strategies.

Construction Strategy	$AP_{BEV\|R40}$			$AP_{3D\|R40}$		
	Easy	Mod.	Hard	Easy	Mod.	Hard
Cross-Attention	41.33	31.50	27.98	33.36	24.46	21.95
Nearest Prototype	41.70	31.12	27.78	33.38	**24.61**	**22.12**
Top-k Soft Fusion	**42.52**	**31.62**	**28.12**	**33.56**	24.57	22.09

5 Conclusion

In this paper, we revisit a key limitation of existing denoising-based monocular 3D detectors, namely the insufficient exploitation of instance-level visual information during denoising learning. To address this issue, we propose ProtoDN, an appearance-aware denoising framework with dynamic visual prototypes. We first construct a dynamically updated visual prototype memory to maintain stable and representative visual priors throughout training. Building on this memory, Geometry–Visual Joint Difficulty Modeling incorporates visual cues into difficulty estimation, while Prototype-Guided Visual Enhancement injects appearance priors into denoising representations, jointly providing richer and more effective visual guidance for denoising. Extensive experiments demonstrate that ProtoDN consistently improves detection performance, with particularly notable gains on challenging instances. Comprehensive ablation studies further validate the effectiveness of each proposed component, highlighting their contribution in advancing the 3D detection task.

AI USE STATEMENT

In this work, large language models (LLMs) were used solely to assist with language polishing and translation during manuscript preparation. They were not used for research design, experimentation, data analysis, or scientific conclusions.

REFERENCES

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *European conference on computer vision*, pp. 139–156. Springer, 2018.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Jiawei Chen, Qi Song, Wenzhong Guo, and Rui Huang. Dsc3d: Deformable sampling constraints in stereo 3d object detection for autonomous driving. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(3):2794–2805, 2024.

Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Andrew G Berneshawi, Huimin Ma, Sanja Fidler, and Raquel Urtasun. 3d object proposals for accurate object class detection. *Advances in neural information processing systems*, 28, 2015.

Yilun Chen, Shijia Huang, Shu Liu, Bei Yu, and Jiaya Jia. Dsgn++: Exploiting visual-spatial relation for stereo-based 3d detectors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4416–4429, 2022.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3354–3361. IEEE, 2012.

John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.

Kuan-Chih Huang, Tsung-Han Wu, Hung-Ting Su, and Winston H Hsu. Monodtr: Monocular 3d object detection with depth-aware transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4012–4021, 2022.

Xueying Jiang, Sheng Jin, Xiaoqin Zhang, Ling Shao, and Shijian Lu. Monomae: Enhancing monocular 3d detection through depth-aware masked autoencoders. *Advances in Neural Information Processing Systems*, 37:11392–11411, 2024.

Junyoung Jung, Seokwon Kim, and Jung Uk Kim. Monosaod: Monocular 3d object detection with sparsely annotated label. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4718–4727, June 2026.

Abhinav Kumar, Yuliang Guo, Zhihao Zhang, Xinyu Huang, Liu Ren, and Xiaoming Liu. Charm3r: Towards unseen camera height robust monocular 3d detector. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8777–8788. IEEE, 2025.

Soyul Lee, Seungmin Baek, and Dongbo Min. Difficulty-aware label-guided denoising for monocular 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5918–5926, 2026.

Changcai Li, Zonghua Gu, Gang Chen, Libo Huang, Wei Zhang, and Huihui Zhou. Real-time stereo-based 3d object detection for streaming perception. *Advances in Neural Information Processing Systems*, 37:115468–115490, 2024.

Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13619–13627, 2022a.

Zhuoling Li, Zhan Qu, Yang Zhou, Jianzhuang Liu, Haoqian Wang, and Lihui Jiang. Diversity matters: Fully exploiting depth clues for reliable monocular 3d object detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2781–2790. IEEE, 2022b.

Shuai Liu, Mingyue Cui, Boyang Li, Quanmin Liang, Tinghe Hong, Kai Huang, and Yunxiao Shan. Fshnet: Fully sparse hybrid network for 3d object detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8900–8909. IEEE, 2025.

Zechen Liu, Zizhang Wu, and Roland Tóth. Smoke: Single-stage monocular 3d object detection via keypoint estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4289–4298. IEEE, 2020.

Zhe Liu, Jinghua Hou, Xiaoqing Ye, Tong Wang, Jingdong Wang, and Xiang Bai. Seed: A simple and effective 3d detr in point clouds. In *European Conference on Computer Vision*, pp. 110–126. Springer, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Yan Lu, Xinzhu Ma, Lei Yang, Tianzhu Zhang, Yating Liu, Qi Chu, Junjie Yan, and Wanli Ouyang. Geometry uncertainty projection network for monocular 3d object detection. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3091–3101. IEEE, 2021.

Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Jana Kosecka. 3d bounding box estimation using deep learning and geometry. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 7074–7082, 2017.

Andretti Naiden, Vlad Paunescu, Gyeongmo Kim, ByeongMoon Jeon, and Marius Leordeanu. Shift r-cnn: Deep monocular 3d object detection with closed-form geometric constraints. In *2019 IEEE international conference on image processing (ICIP)*, pp. 61–65. IEEE, 2019.

Liang Peng, Xiaopei Wu, Zheng Yang, Haifeng Liu, and Deng Cai. Did-m3d: Decoupling instance depth for monocular 3d object detection. In *European Conference on Computer Vision*, pp. 71–88. Springer, 2022.

Liang Peng, Junkai Xu, Haoran Cheng, Zheng Yang, Xiaopei Wu, Wei Qian, Wenxiao Wang, Boxi Wu, and Deng Cai. Learning occupancy for monocular 3d object detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10281–10292. IEEE, 2024.

Fanqi Pu, Yifan Wang, Jiru Deng, and Wenming Yang. Monodgp: Monocular 3d object detection with decoupled-query and geometry-error priors. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6520–6530. IEEE, 2025.

Shaoshuai Shi, Li Jiang, Jiajun Deng, Zhe Wang, Chaoxu Guo, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn++: Point-voxel feature set abstraction with local vector representation for 3d object detection. *International Journal of Computer Vision*, 131(2):531–551, 2023.

Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. Disentangling monocular 3d object detection. In *2019 IEEE/CVF international conference on computer vision (ICCV)*, pp. 1991–1999. IEEE, 2019.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Yongzhi Su, Yan Di, Guangyao Zhai, Fabian Manhardt, Jason Rambach, Benjamin Busam, Didier Stricker, and Federico Tombari. Opa-3d: Occlusion-aware pixel-wise aggregation for monocular 3d object detection. *IEEE Robotics and Automation Letters*, 8(3):1327–1334, 2023.

Kiet Dang Vu, Trung Thai Tran, and Duc Dung Nguyen. Monovqd: Monocular 3d object detection with variational query denoising and self-distillation. *arXiv preprint arXiv:2506.14835*, 2025.

Kiet Dang Vu, Trung Thai Tran, Kien Nguyen Do Trung, and Duc Dung Nguyen. Mono3dv: Monocular 3d object detection with 3d-aware bipartite matching and variational query denoising. *arXiv preprint arXiv:2601.01036*, 2026.

Yifan Wang, Yian Zhao, Fanqi Pu, Xiaochen Yang, Yang Tang, Xi Chen, and Wenming Yang. Span: spatial-projection alignment for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40762–40771, 2026.

Hai Wu, Chenglu Wen, Shaoshuai Shi, Xin Li, and Cheng Wang. Virtual sparse convolution for multimodal 3d object detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21653–21662. IEEE, 2023a.

Wenhao Wu, Hau San Wong, and Si Wu. Semi-supervised stereo-based 3d object detection via cross-view consensus. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17471–17481, 2023b.

Zizhang Wu, Yuanzhu Gan, Lei Wang, Guilian Chen, and Jian Pu. Monopgc: Monocular 3d object detection with pixel geometry contexts. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4842–4849. IEEE, 2023c.

Zizhang Wu, Yuanzhu Gan, Yunzhe Wu, Ruihao Wang, Xiaoquan Wang, and Jian Pu. Fd3d: Exploiting foreground depth map for feature-supervised monocular 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6189–6197, 2024.

Longfei Yan, Pei Yan, Shengzhou Xiong, Xuanyu Xiang, and Yihua Tan. Monocd: Monocular 3d object detection with complementary depths. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10248–10257. IEEE, 2024.

Sunghun Yang, Minhyeok Lee, Jungho Lee, and Sangyoun Lee. Monoclue: Object-aware clustering enhances monocular 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11721–11729, 2026.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *International Conference on Learning Representations*, 2023a.

Renrui Zhang, Han Qiu, Tai Wang, Ziyu Guo, Ziteng Cui, Yu Qiao, Hongsheng Li, and Peng Gao. Monodetr: Depth-guided transformer for monocular 3d object detection. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9121–9132. IEEE, 2023b.

Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects are different: Flexible monocular 3d object detection. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 3288–3297. IEEE, 2021.

Zhihao Zhang, Abhinav Kumar, Girish Chandar Ganesan, and Xiaoming Liu. Unleashing the power of chain-of-prediction for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18755–18765, 2026.

A Additional Implementation Details

Network Architecture. Following the overall detector architecture described in the main paper, the multi-scale image features are first processed by a 2D Transformer composed of three encoder layers and three decoder layers. The resulting representations are then passed to a three-layer depth-aware 3D decoder for 3D attribute modeling. All Transformer modules use a hidden dimension of 256 with eight attention heads, while deformable attention employs four sampling points at each feature level. Iterative box refinement is enabled during decoding to progressively refine the predicted boxes. For depth modeling, we adopt the linear-increasing discretization (LID) scheme with 80 depth bins covering the range from 10^{-3} m to 60 m. Following the query configuration of the underlying framework, 50 detection queries are used during inference, while 11 query groups are employed during training.

Data Processing and Augmentation. During training, we apply several data augmentation strategies, including photometric distortion, random horizontal flipping, random cropping, and 3D MixUp. The probabilities of random flipping, random cropping, and 3D MixUp are all set to 0.5. Random cropping is further combined with scale and image-center perturbations, both with a magnitude of 0.05. After affine transformation, all input images are resized to a fixed resolution and normalized using the ImageNet mean and standard deviation. Unless otherwise specified, all experiments use a fixed random seed of 444 for reproducibility.

Denoising Configuration. During training, we construct seven denoising groups, each containing positive and negative noisy copies of the ground-truth instances. The base perturbation scales for the projected box and depth error are set to 0.4 and 0.8, respectively, while the label-noise ratio is set to 0.5. These base perturbations are further adaptively modulated according to the instance-level difficulty described in the main paper. The resulting denoising queries are used only during training and are removed entirely at inference.

Visual Prototype Implementation. The instance features used for visual prototype construction are first projected into a 256-dimensional embedding space and ℓ_2-normalized. Cosine similarity is then used for prototype assignment and visual-certainty estimation. During training, the visual prototypes are updated every five iterations, with $\epsilon = 10^{-6}$ used to avoid unstable feature aggregation when the assignment weight of a prototype becomes too small. The visual prototype memory and all prototype-related operations are enabled only during training, introducing no additional inference cost.

Training Objective and Matching. The detailed loss weights and bipartite matching costs used in our experiments are summarized in Table 6. The overall training objective consists of classification, 2D localization, 3D attribute, depth, and region losses. During training, Hungarian bipartite matching is first performed to establish one-to-one correspondences between prediction queries and ground-truth targets, using a matching cost composed of classification, 2D box, GIoU, and 3D center terms. The detection losses are then computed according to the resulting assignments, with their corresponding weights listed in the table.

B More Qualitative Results

False Positive Analysis. We conduct a qualitative analysis of false-positive detections by comparing ProtoDN with representative baselines across diverse driving scenes on the KITTI validation set, as shown in Fig. 4. The orange dashed ellipses highlight false-positive predictions that do not correspond to any ground-truth object. MonoDGP and MonoDLGD produce several spurious detections in background regions, whereas ProtoDN substantially suppresses these responses and yields predictions that are more consistent with the ground truth. In particular, our method exhibits fewer isolated boxes in visually ambiguous regions and maintains cleaner predictions around valid objects, indicating a better separation between target-related and background features. This behavior is consistent with our appearance-aware denoising design, where instance-level visual cues provide additional guidance beyond geometric perturbations. As a result, the model is less prone to generating unreliable predictions from background patterns and shows improved robustness in complex driving scenes.

Table 6: Loss weights and matching costs used in our experiments.

Type	Term	Weight
Training Loss	Classification ($\mathcal{L}_{\text{cls}}$)	2.0
	2D Box L1 ($\mathcal{L}_{\text{bbox}}$)	5.0
	GIoU ($\mathcal{L}_{\text{giou}}$)	2.0
	3D Center ($\mathcal{L}_{\text{center}}$)	10.0
	Dimension ($\mathcal{L}_{\text{dim}}$)	1.0
	Orientation ($\mathcal{L}_{\text{angle}}$)	1.0
	Depth ($\mathcal{L}_{\text{depth}}$)	1.0
	Depth Map ($\mathcal{L}_{\text{depthmap}}$)	1.0
	Region ($\mathcal{L}_{\text{region}}$)	1.0
Matching Cost	Classification	2.0
	2D Box	5.0
	GIoU	2.0
	3D Center	10.0



(a) MonoDGP **(b) MonoDLGD** **(c) ProtoDN**

Figure 4: Qualitative comparison of false-positive detections on the KITTI validation set.

Localization Quality Analysis. We further conduct a qualitative analysis of 3D localization quality by comparing ProtoDN with MonoDGP and MonoDLGD across several KITTI validation scenes, as shown in Fig. 5. Compared with the two baselines, the predicted 3D boxes of ProtoDN are more closely aligned with the ground-truth boxes in both the camera and bird's-eye views. The improvement is reflected not only in the box position, but also in the orientation and spatial extent, while the baseline predictions exhibit noticeable heading deviations and partial box misalignment for several objects. In the bird's-eye view, ProtoDN also shows more consistent overlap with the ground truth, suggesting more accurate center localization and geometric estimation. These observations indicate that introducing instance-level visual information into denoising provides complementary appearance guidance for 3D reconstruction, helping the model produce more geometrically consistent predictions, especially for challenging instances where geometry-only cues may be insufficient.



(a) MonoDGP (b) MonoDLGD (c) ProtoDN

Figure 5: Qualitative comparison of 3D bounding box localization on the KITTI validation set.

Table 7: Ablation study on the number of class-specific visual prototypes.

K	$AP_{BEV\|R40}$			$AP_{3D\|R40}$		
	Easy	Mod.	Hard	Easy	Mod.	Hard
3	41.42	30.57	27.74	33.45	24.54	21.14
5	**42.52**	**31.62**	**28.12**	**33.56**	**24.57**	**22.09**
7	42.43	31.15	27.87	32.94	24.08	21.82

C Additional Ablation Studies

Number of Class-specific Visual Prototypes. We investigate the influence of the number of class-specific visual prototypes K in Table 7. Using only three prototypes provides limited capacity to capture diverse appearance patterns, particularly on challenging instances, whereas increasing K from 3 to 5 improves Hard $AP_{3D|R40}$ by 0.95% and Moderate $AP_{BEV|R40}$ by 1.05%. Further increasing the number of prototypes to 7 does not bring additional benefits; instead, Moderate $AP_{3D|R40}$ decreases by 0.49% compared with $K = 5$. Overall, $K = 5$ achieves the best performance across all difficulty levels for both BEV and 3D detection, suggesting a suitable balance between prototype diversity and representation compactness. We therefore use $K = 5$ in all experiments.

Prototype Update Strategy. We further investigate the update strategy of the dynamic visual prototype memory. In the *hard update* variant, each ground-truth RoI feature is assigned only to its most similar prototype within the corresponding class. Features assigned to the same prototype are averaged and then used to update that prototype through EMA, while prototypes receiving no assignments remain unchanged. In contrast, our *soft update* computes similarity-based softmax weights over all class-specific prototypes, allowing each instance feature to contribute to multiple prototypes with different weights. The weighted features are aggregated within the current batch and subsequently incorporated into the prototype memory through EMA. All other settings are kept unchanged.

As shown in Table 8, soft updating consistently outperforms hard updating across all reported metrics. In particular, it improves Easy and Hard $AP_{3D|R40}$ by 1.94% and 1.45%, respectively, while

Table 8: Ablation study on different prototype update strategies.

Update Strategy	$AP_{BEV\|R40}$			$AP_{3D\|R40}$		
	Easy	Mod.	Hard	Easy	Mod.	Hard
Hard Update	40.85	30.13	27.39	31.62	23.89	20.64
Soft Update (Ours)	**42.52**	**31.62**	**28.12**	**33.56**	**24.57**	**22.09**

Table 9: Ablation study on different geometry–visual certainty fusion strategies.

Fusion Strategy	$AP_{BEV\|R40}$			$AP_{3D\|R40}$		
	Easy	Mod.	Hard	Easy	Mod.	Hard
Additive Fusion	40.83	30.02	26.37	32.08	24.11	20.85
Multiplicative Fusion (Ours)	**42.52**	**31.62**	**28.12**	**33.56**	**24.57**	**22.09**

Moderate $AP_{BEV|R40}$ increases by 1.49%. These results suggest that the winner-take-all assignment in hard updating may lead to less stable prototype adaptation, whereas similarity-weighted aggregation allows the prototype memory to evolve more smoothly and better preserve diverse instance-level appearance information. We therefore adopt the soft update strategy in our final model.

Geometry–Visual Certainty Fusion. We also examine how geometric and visual certainties should be combined for difficulty-aware perturbation. Our default *multiplicative fusion* progressively introduces visual certainty after the warm-up stage. As an alternative, we construct an *additive fusion* variant that smoothly transitions from geometric certainty to a weighted combination:

$$c^a_{\text{joint},i} = (1-\lambda_t)c^a_{\text{geo},i} + \lambda_t \left(0.7c^a_{\text{geo},i} + 0.3c_{\text{vis},i}\right), \qquad a \in \{v, d\}. \tag{15}$$

The resulting certainty is used in the same manner to modulate the perturbation magnitudes of the projected box and depth, with all other settings unchanged. Table 9 shows that multiplicative fusion consistently performs better than additive fusion across all difficulty levels. The improvements are particularly evident on challenging instances, with gains of 1.75% in Hard $AP_{BEV|R40}$ and 1.24% in Hard $AP_{3D|R40}$. Moderate $AP_{BEV|R40}$ also increases by 1.60%. Unlike additive fusion, where a high certainty from one modality can partially compensate for a low certainty from the other, multiplicative fusion jointly constrains the two cues and reduces the final certainty whenever either geometric or visual reliability is low. The results indicate that such joint modulation provides a more suitable difficulty estimate for controlling denoising perturbations.