MonoDLGD: Difficulty-Aware Label-Guided Denoising for Monocular 3D Object Detection

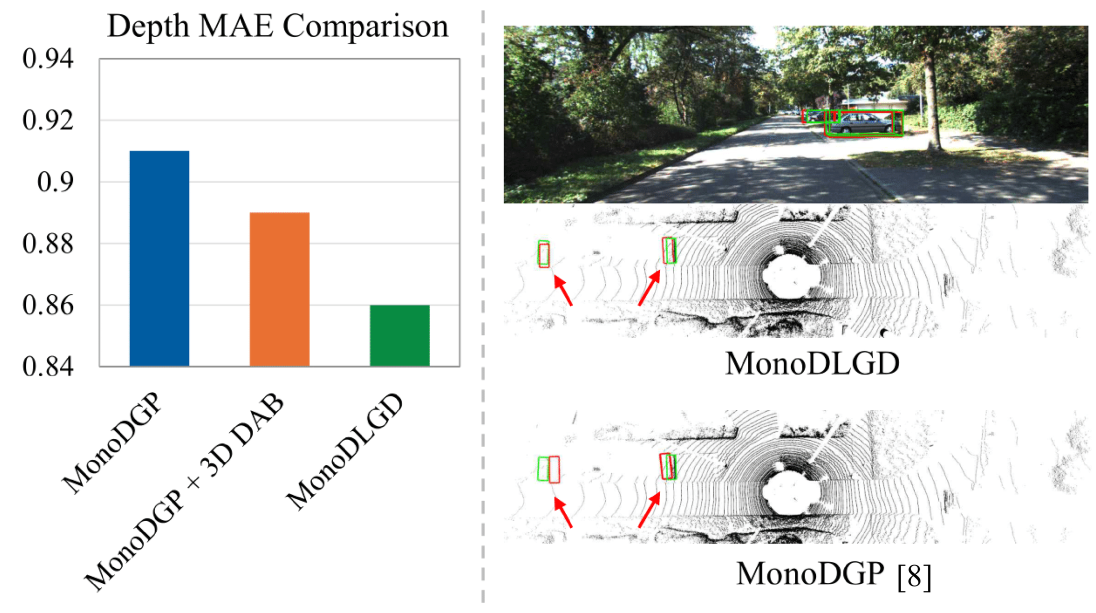
Soyul Lee*, Seungmin Baek*, Dongbo Min†
Ewha Womans University, Seoul, Korea
* Equal, †corresponding author

AAAI-26 / IAAI-26 / EAAI-26

Motivation

◆ Problem

- Monocular 3D object detection is inherently ill-posed due to the lack of explicit depth cues.
- **Localization Error**: Auxiliary single-image depth estimation has been introduced to mitigate this issue.
  → yet it remains insufficient to resolve 3D localization errors.
- **Overlooked Difficulty**: Object detection difficulty in monocular settings is inherently multi-factor (scale, distance, truncation, occlusion).
  ⇨ Ignoring instance level difficulty degrades stability & representation quality.



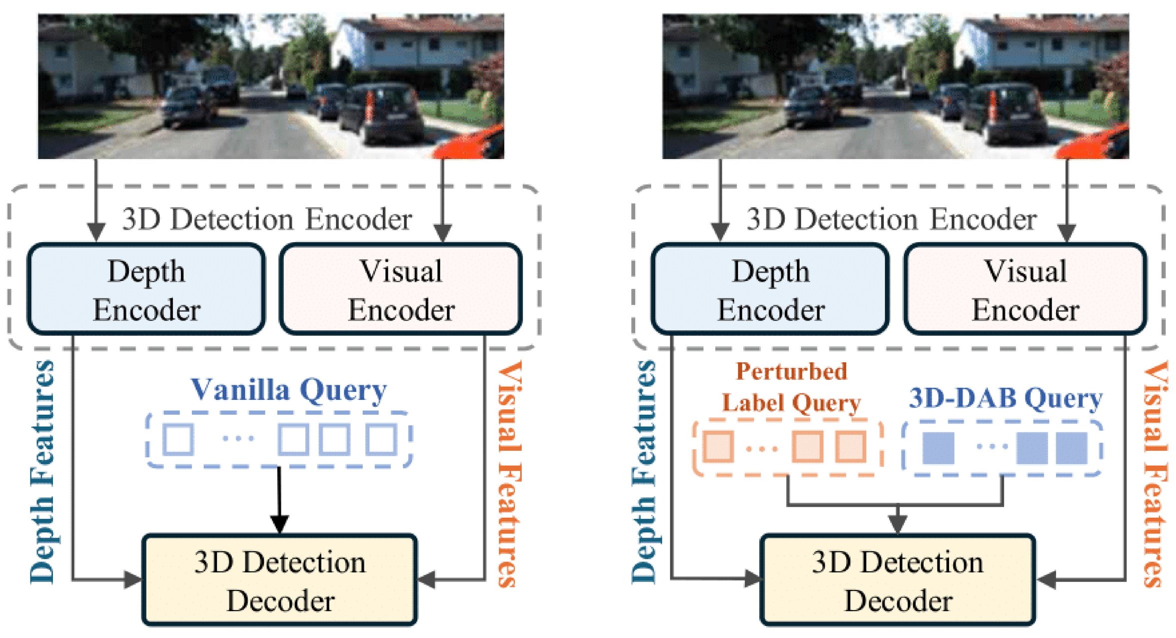
◆ Our Goal

- Learn **robust geometric representations** for objects by **explicitly modeling instance-level detection difficulty and providing strong geometry supervision.**

◆ Key Idea

- **Difficulty-Adaptive Denoising**: Inject difficulty-adaptive perturbations into 3D ground-truth(GT) labels based on the predicted uncertainty and to learn reconstruct them during training.

⇨ **Explicit geometric supervision.**
⇨ **Enable robust & geometry-aware representation learning.**



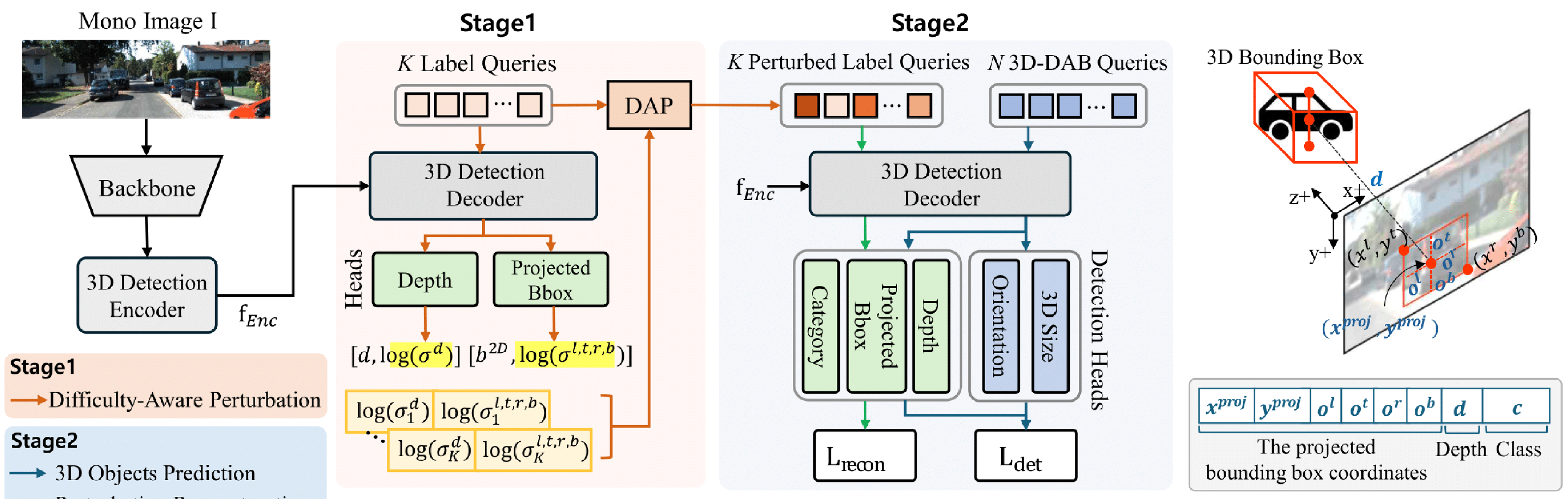
(a) MonoDGP (b) MonoDLGD (Ours)

◆ Contributions

- MonoDLGD introduces **label perturbation and reconstruction** guided by prediction uncertainty, effectively leveraging **explicit geometry supervision.**
- Demonstrates that **modeling instance-level uncertainty** significantly enhances monocular 3d detection accuracy.
- Achieve SOTA on KITTI benchmark without any additional inference overhead.

Method

◆ Overview



K: Number of Objects in a mini-batch N: Number of Anchor Queries in a mini-batch

➢ 3D-Dynamic Anchor Boxes (3D-DAB):

- Embed spatial priors (projected bboxes, depths) into queries to align with perturbed label features.
- **Query Formulation:** b^{proj}(projected bbox), d (depth), c (class)

$$q_i = [b_i^{proj}, d_i, c_i] \in \mathbb{R}^{7+C} \qquad b^{proj} = [x^{proj}, y^{proj}, o^l, o^t, o^r, o^b] \in \mathbb{R}^6$$

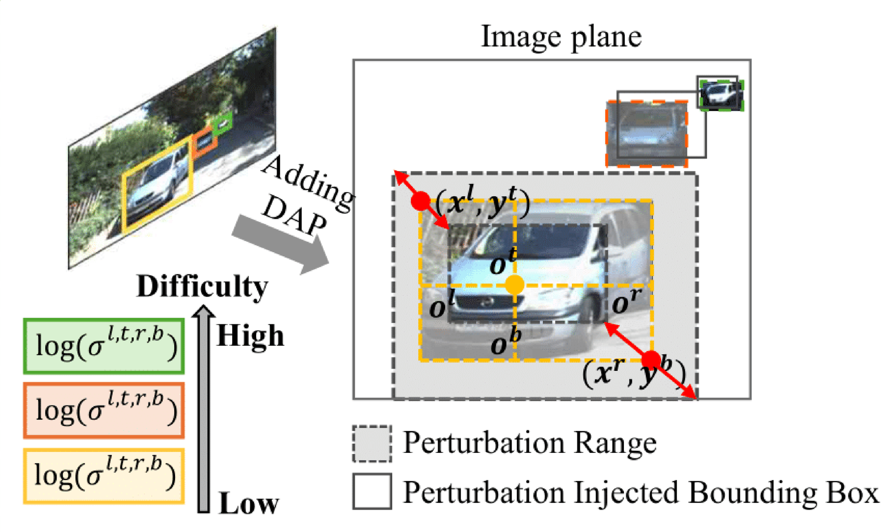
- By utilizing b^{proj} and d, the model constraints the search space to geometrically meaningful regions.

➢ Stage 1: Difficulty-Aware Perturbation (DAP):

- DAP adaptively scales perturbation strength based on estimated instance-level detection difficulty
- **Difficulty-Scoring:** Estimates aleatoric uncertainty $\log(\sigma^v)$ as a difficulty proxy to compute a normalized difficulty score $\widehat{c^v} \in$ **[0, 1]**.

$$c^v = \exp(-\log(\sigma^v)), \quad v \in \{d, l, t, r, b\}. \quad \hat{c}^v = \frac{c^v - c^v_{\min}}{c^v_{\max} - c^v_{\min}}$$

- **Adaptive Strategy:**
  Hard instances (High σ^v)
  → Weaker perturbations (preserve geometry)
  Easy instances (Low σ^v)
  → Stronger perturbations (regularization)



DAP for projected bounding box.

➢ Stage2: Difficulty-Aware Reconstruction

- The shared decoder simultaneously performs 3D object prediction and label reconstruction to provide explicit geometric supervision.
- Uncertainty-adaptive training → Employ the **Laplacian aleatoric uncertainty loss**

Depth reconstruction loss:

$$L^d_{recon} = \sum_{i=1}^{K} \left(\frac{\sqrt{2}}{\sigma_i^d} ||d_{gt,i} - d_{recon,i}||_1 + log(\sigma_i^d) \right)$$

Bbox reconstruction loss:

$$L^{bbox}_{recon} = \sum_{i=1}^{K} \left(\sum_{v \in \{l,r\}} \left(\frac{\sqrt{2}}{\sigma_i^v} \left\| x^v_{gt,i} - x^v_{recon,i} \right\|_1 + \log(\sigma_i^v) \right) + \sum_{v \in \{t,b\}} \left(\frac{\sqrt{2}}{\sigma_i^v} \left\| y^v_{gt,i} - y^v_{recon,i} \right\|_1 + \log(\sigma_i^v) \right) \right)$$

Total label reconstruction loss:

$$L_{recon} = \lambda_{bbox} L^{bbox}_{recon} + \lambda_d L^d_{recon} + \lambda_{cls} \boxed{L^{cls}_{recon}}$$

Class reconstruction loss

Total loss:

$$L = L_{recon} + \boxed{L_{det}}$$

Detection loss

Results

◆ Quantitative Results

Table 2: **Comparisons on the KITTI test and validation sets (Car category)**. We **bold** the best results and underline the second-best results.

Methods	Extra data	Reference	Test, $AP_{BEV\|R40}$ Easy	Mod.	Hard	Test, $AP_{3D\|R40}$ Easy	Mod.	Hard	Val, $AP_{BEV\|R40}$ Easy	Mod.	Hard	Val, $AP_{3D\|R40}$ 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		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		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	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		RAL 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
DEVIANT [1]	None	ECCV 2022	29.65	20.44	17.43	21.88	14.46	11.89	32.60	23.04	19.99	24.63	16.54	14.52
MonoDDE [2]		CVPR 2022	33.58	23.46	20.37	24.93	17.14	15.10	35.51	26.48	23.07	26.66	19.75	16.72
MonoUNI [3]		NeurIPS 2023	-	-	-	24.75	16.73	13.49	-	-	-	24.51	17.18	14.01
MonoDETR [4]		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
MonoCD [5]		CVPR 2024	33.41	22.81	19.57	25.53	16.59	14.53	34.60	24.96	21.51	26.45	19.37	16.38
FD3D [6]		AAAI 2024	34.20	23.72	20.76	25.38	17.12	14.50	36.98	26.77	23.16	28.22	20.23	17.04
MonoMAE [7]		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 [8]		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
Ours	None		**36.63**	**25.3**	**23.13**	**29.11**	**19.87**	**17.74**	**41.68**	**30.53**	**27.76**	**34.89**	**25.19**	**21.78**
Improvement over Second-Best Method	-		+1.39	+0.07	+1.11	+2.76	+1.03	+0.96	+1.42	+2.33	+3.34	+4.13	+2.85	+2.76
Improvement over MonoDGP Baseline	-		+1.39	+0.07	+1.11	+2.76	+1.15	+1.77	+2.28	+2.33	+3.34	+4.13	+2.85	+2.76

◆ Ablation Study

Idx	3D-DAB	Perturb.	L^{bbox}_{recon}	Val, $AP_{BEV\|R40}$ Easy	Mod.	Hard	Val, $AP_{3D\|R40}$ Easy	Mod.	Hard
(a)	×	×	×	39.40	28.20	24.42	30.76	22.34	19.02
(b)	O	×	×	36.85	26.72	23.21	27.82	20.64	17.78
(c)	O	UN	L1 Loss	40.32	30.13	26.53	31.99	23.82	20.65
(d)	O	UN	LU Loss	41.16	30.31	26.54	33.82	24.7	21.19
(e): Ours	O	DAP	LU Loss	41.68	30.53	27.76	34.89	25.19	21.78

UN: Uniform Noise
LU: Laplacian Uncertainty

◆ Computational Cost

Method	Val, $AP_{BEV\|R40}$ Easy	Mod.	Hard	Val, $AP_{3D\|R40}$ Easy	Mod.	Hard	GFLOPs↓	Time (ms)↓
MonoDETR* [4]	36.38	26.19	22.29	27.34	19.33	16.04	59.7	35.2
+Ours	38.59	27.65	23.62	29.79	21.63	18.17	59.8	35.5
MonoDGP [8]	39.40	28.20	24.42	30.76	22.34	19.02	69.0	42.4
+Ours	41.68	30.53	27.76	34.89	25.19	21.78	69.3	42.7

Reference

[1] "Deviant: Depth equivariant network for monocular 3d object detection." *ECCV*, 2022.
[2] "Diversity matters: Fully exploiting depth clues for reliable monocular 3d object detection." *CVPR*. 2022.
[3] "Monouni: A unified vehicle and infrastructure-side monocular 3d object detection network with sufficient depth clues." *NIPS.* 2023.
[4] "MonoDETR: Depth-guided transformer for monocular 3D object detection." ICCV 2023.
[5] "Monocd: Monocular 3d object detection with complementary depths." CVPR 2024.
[6] "Fd3d: Exploiting foreground depth map for feature-supervised monocular 3d object detection." AAAI 2024.
[7] "MonoMAE: Enhancing monocular 3D detection through depth-aware masked autoencoders." NIPS 2024.
[8] "Monodgp: Monocular 3d object detection with decoupled-query and geometry-error priors." CVPR. 2025.