MonoPure: Multi-Component Purification via Disentangled, Projective Representations for Monocular 3D Object Detection

Yeon Woo Cho , **Jung Woo Cheon** , **Seung-hyeok Back** and **Seok Bong Yoo***
Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju, Korea
sbyoo@jnu.ac.kr

Abstract

Monocular 3D object detection is a cost-efficient alternative to multisensor systems, yet it remains fragile to multi-component adversarial attacks that perturb the image and tamper with camera calibration. Compounded distortions degrade 3D reasoning by disrupting the correspondence between the 3D geometry and 2D image plane. To address this problem, this work proposes MonoPure, a monocular 3D object detection framework that performs multi-component purification via disentangled and projective representations. MonoPure incorporates a disentangled purification and segmentation module that purifies the image data, with a target-region probability map steering diffusion-based purification to focus on task-relevant regions. In addition, MonoPure presents a 3D detection decoder that integrates 2D skeleton keypoints as object-level spatial cues, enabling occlusion-robust 3D detection. Finally, a projective calib-purification module restores compromised intrinsics by iteratively minimizing the reprojection error between projected 3D boxes and calibration-invariant 2D detection boxes. The experiments confirm that MonoPure outperforms prior detectors under multi-component attacks and occlusion.

1 Introduction

Monocular 3D object detection has gained significant attention in autonomous perception as a cost-effective method for reconstructing 3D environments from single red, green, and blue (RGB) images by leveraging appearance and geometric cues. However, the inherent challenge of recovering 3D depth from a 2D plane renders these models exceptionally sensitive to subtle feature distortions. This sensitivity makes them highly vulnerable to adversarial perturbations injected through practical channels. As illustrated in Figure 1(a), the security of monocular 3D detection systems is threatened by practical adversarial scenarios targeting both sensory data and system configurations. On-board tampering executed via malware, memory corruption, or physical interception allows an adversary to compromise the integrity of various internal components beyond simple image-level perturbations. Furthermore, transmission-level attacks exploit RSU or server communications to inject malicious data or conduct replay attacks by retransmitting outdated system metadata or environmental parameters.

Based on the 3D detection performance under the attack scenarios shown in Figure 1(b), state-of-the-art models such as MonoDGP [Pu *et al.*, 2025] and MonoDETR [Zhang *et al.*, 2023] are exceptionally susceptible to these adversarial threats. Image adversarial attacks pose a severe threat by substantially degrading detection accuracy. However, the impact becomes far more destructive under multi-component attacks, where detection performance completely fails, reaching near-zero. Such extreme sensitivity to appearance and geometric mismatches necessitates monocular 3D object detectors that incorporate robust purification methods to ensure reliable 3D reasoning under adversarial conditions.

Input-level purification provides a practical defense but shifts distribution and oversmoothes geometric cues. This hampers localization under occlusion where evidence is weak. Moreover, it fails against multi-component attacks corrupting calibration. Visual restoration alone ignores geometric inconsistencies, leading to catastrophic 3D failures. To address these problems, this work presents MonoPure, a framework for multi-component purification via disentangled and projective representations for monocular 3D object detection.

Subsequently, the keypoint detector employs a curriculum strategy that adapts from labeled data to unlabeled scenarios via masked self-training. A dual-graph network refines these predictions by enforcing intra-instance geometry and inter-instance context to mitigate occlusion. These robust keypoints are integrated into the 3D decoder via depth-KP cross-attention, serving as spatial anchors for geometric reasoning. Ultimately, fusing these structural priors with visual and depth features stabilizes 3D localization against adversarial attacks and severe occlusion. Finally, this approach exploits the observation that 2D boxes predicted by the 2D detection decoder remain largely insensitive to camera calibration attacks. Treating these boxes as reliable references, the method purifies the compromised calibration by iteratively minimizing the spatial gap between the 2D boxes reprojected from the 3D predictions and the reliable 2D detection anchors. In Figure 1(c), MonoPure remains robust under image

* Corresponding author

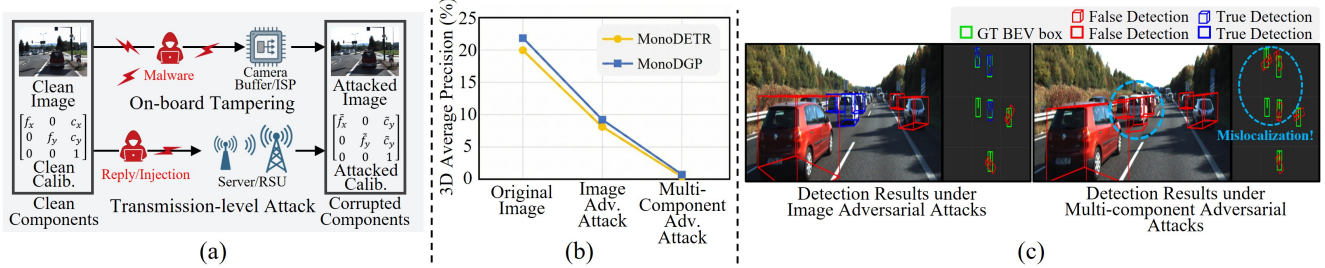

Figure 1: (a) Practical adversarial attack scenarios in autonomous driving. (b) Image-only and image and calibration attacks on monocular 3D object detection. (c) Visualization of 3D box failures under image-only and image and calibration attacks.

and multi-component adversarial attacks. Our source code and appendix are available at https://github.com/yeonwoo29/MonoPure. We summarize the contributions below:

- This work presents a monocular 3D object detection framework that, for the first time, explicitly addresses multi-component adversarial attacks that jointly corrupt input images and camera calibration parameters.
- This paper introduces a disentangled purification and segmentation module that separates backbone features into two complementary subspaces for purification and segmentation. This design removes adversarial perturbations while preserving detection-critical structures.
- This approach presents a 3D decoder with 2D skeleton keypoint detector that combines mask-enhanced self-curriculum learning with a dual-graph 2D keypoint refinement network. This module localizes and reconstructs the keypoints under occlusion.
- This paper introduces a projective calib-purification module that iteratively narrows the spatial gap between 3D reprojections and 2D anchors to refine camera calibration parameters, ensuring robust 3D localization.

2 Related Work

2.1 Adversarial Attack

Vision-based perception systems for autonomous driving are highly vulnerable to adversarial attacks [Liu and Jiang, 2025; Zhu *et al.*, 2023; Kou *et al.*, 2023], and the attack surface is rapidly expanding. Beyond image-only perturbations [Zhang *et al.*, 2025; Chen *et al.*, 2024], recent studies show that autonomous-driving perception can be attacked by corrupting sensory inputs and system parameters [Cheng *et al.*, 2022]. Yet, despite the safety-critical role of monocular 3D object detection, adversarial robustness remains underexplored in this domain. Existing efforts mostly focus on single-component attacks that target either the input image [Li *et al.*, 2023] or camera calibration parameters [Zhou *et al.*, 2021]. In practice, such access is often realistic, as both sensory inputs and system parameters are exposed. Unlike single-component attacks, multi-component attacks couple image perturbations with calibration corruption, causing severe projection inconsistencies and making 3D reasoning highly unstable. Therefore, MonoPure is designed to explicitly target multi-component threats by unifying the handling of image-level and system-level corruption, enabling robust monocular 3D detection.

2.2 Adversarial Purification

Adversarial attacks pose a serious threat to vision-based perception systems, prompting extensive research on input-level purification as a practical defense. Prior work has explored learning-based purification, including autoencoder-style reconstruction [Niu and Yang, 2023] and diffusion-based denoising [Yoon *et al.*, 2021; Kang *et al.*, 2024; Choi *et al.*, 2024]. However, most diffusion-based purification methods have been developed for image classification [Lei *et al.*, 2025] and segmentation [Lee *et al.*, 2024], where preserving global semantics is often sufficient. In monocular 3D detection, generic purification can blur object-level geometric cues and destabilize 3D localization. To address this, MonoPure uses target region probability maps for object-centric purification and disentangles reconstruction and segmentation features to enable task-aware denoising that preserves geometry-critical cues. Moreover, monocular 3D object detection is vulnerable to calibration attacks that perturb camera intrinsics and distort 3D geometry, but existing purification methods do not handle attacked calibration parameters. MonoPure purifies both images and calibration. In addition, since real-time inference is critical in this field, MonoPure enables fast purification compared to iterative diffusion-based defenses.

2.3 Monocular 3D Object Detection

Monocular 3D object detection [Liu *et al.*, 2025] has attracted considerable attention because it is cost-efficient and simple to deploy compared to LiDAR- [Gambashidze *et al.*, 2024] or multi-view [Xue *et al.*, 2025; Qi *et al.*, 2025] systems. Early approaches [Chen *et al.*, 2016] impose geometric constraints by aligning 2D detections with projected 3D bounding boxes. More recent transformer-based methods [Shi *et al.*, 2025] improve depth reasoning and localization by applying the global context and learned geometric priors. However, these detectors remain fragile when input is corrupted, and performance can degrade sharply under adversarial perturbations or occlusion [Lin *et al.*, 2024]. To address this, MonoPure bridges these lines of research by unifying adversarially robust purification and occlusion-robust skeleton keypoint cues. It improves reliability against adversarial perturbations through purification and tackles occlusion using skeleton keypoint cues.

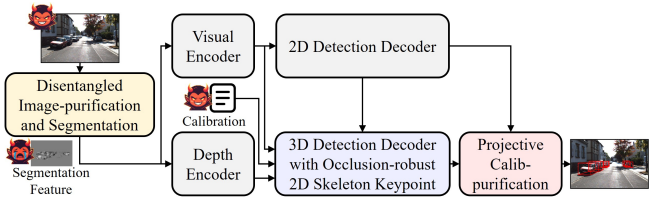

Figure 2: Architecture of the MonoPure framework.

3 Method

3.1 Overview

Figure 2 presents MonoPure, a robust monocular 3D object detection framework designed to withstand multi-component adversarial attacks that jointly corrupt the input image and camera calibration. Given an attacked image, MonoPure employs a disentangled module that separates purification and segmentation representations, using segmentation-aware guidance to suppress adversarial noise while preserving the structure for detection. Following MonoDGP [Pu *et al.*, 2025], the purified features are processed using a visual encoder for 2D localization and a depth encoder for geometric reasoning [Duan *et al.*, 2019]. MonoPure employs a 3D detection decoder with occlusion-robust 2D skeleton keypoints, using object-level spatial cues and refined semantic keypoint guidance to stabilize 3D localization under occlusion. Finally, MonoPure corrects compromised intrinsics at inference by minimizing the reprojection error between 3D-to-2D projections and 2D detection boxes, providing geometrically consistent 3D predictions.

3.2 Disentangled Image-purification and Segmentation

To improve robustness against attacked input, this work introduces a disentangled purification and segmentation module (Figure 3). Conventional purification can often be overly aggressive in denoising, blurring object boundaries and weakening appearance–geometry correlations. Motivated by task-specific disentangling [Yang *et al.*, 2021], we explicitly separate the encoder representation into a purification feature f_p for reconstruction and a segmentation feature f_s for detection-oriented structure learning.

This approach trains a disentangling encoder that splits the shared representation into channel-wise subspaces, using task-driven objectives for each branch and an independence constraint that reduces feature correlation. Given an attacked image, f_p is input into a reconstruction decoder to obtain a restored image. Next, f_p is decoded by the reconstruction decoder to produce an image I_r, and a clean segmentation feature f_s is then extracted from the purified I_r. This procedure directly reconstructs the image from the encoded features, which can substantially reduce the number of costly denoising steps required by the subsequent diffusion purifier.

Target-Probability guided Diffusion Purifier. This work encodes the restored image into the latent space using a variational autoencoder $\mathcal{E}$ [Hinton and Salakhutdinov, 2006], producing a latent code z_0. Next, this work applies the forward

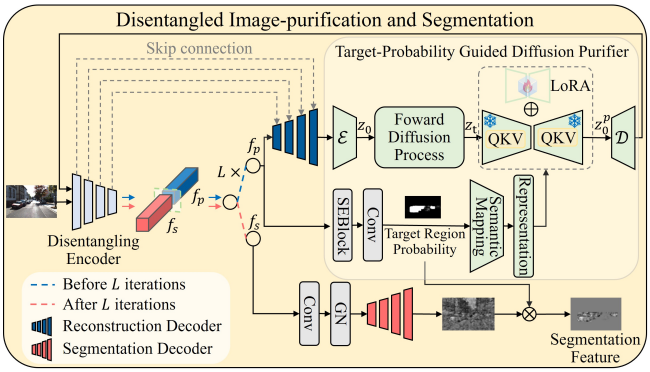

Figure 3: Disentangled image-purification and segmentation.

diffusion process to obtain z_t, which is input into the diffusion denoiser. At this time, f_p is routed to the reconstruction branch and applied to produce a restored image, which is refined using the target-guided diffusion purifier. To generate guidance, this work applies convolution layers followed by a squeeze and excitation (SE) block [Hu *et al.*, 2018] to emphasize the foreground structure and suppress irrelevant background responses. The resulting spatial probability map R, consisting of probability values between 0 and 1, guides the diffusion denoising process by focusing the refinement on detection-relevant regions, and is computed as follows:

$$S = relu\big(W_2 \cdot sigmoid(W_1 \cdot \mathcal{P}(f_p))\big), \tag{1}$$

$$R = relu\big(W_p * (S \odot f_p)\big), \quad R \in [0,1], \tag{2}$$

where $\mathcal{P}(\cdot)$ denotes global average pooling, $\odot$ represents channel-wise multiplication, $*$ indicates a convolution operator, and $\cdot$ signifies matrix multiplication. Moreover, W_1 and W_2 denote learnable convolutional parameters for the gating and spatial projection. z_t and the probability map R are fed into the diffusion denoiser, as formulated below:

$$f_\theta(z_t, c, R, t) = z_t + \frac{z_t - \sqrt{1-\alpha_t}\hat{\epsilon}_\theta(z_t, c, R, t)}{\sqrt{\alpha_t}}, \tag{3}$$

where f_θ is the consistency model parameterized by θ that predicts the clean latent code, and c represents the auxiliary conditioning signals and other contexts. Furthermore, $\hat{\epsilon}_\theta(\cdot)$ indicates the noise prediction of the denoiser, and α_t represents the cumulative noise schedule at timestep t.

To ensure efficiency, this work performs diffusion in the latent space and adopts low-rank adaptation (LoRA) [Hu *et al.*, 2022] to specialize a pretrained Stable Diffusion [Rombach *et al.*, 2022] denoiser for purification with only a few training parameters. Specifically, this work implements LoRA adapters in Stable Diffusion layers and optimizes them using a latent consistency objective. To our knowledge, MonoPure is among the first attempts to apply a target-region probability map as spatial guidance in diffusion-based purification for monocular 3D detection. This approach defines a consistency target z_{target} from a teacher denoising trajectory and trains the student to match it in a single step. The LoRA loss is:

$$\mathcal{L}_{\text{LoRA}}(\theta) = \|f_\theta(z_t, c, R, t) - z_{\text{target}}\|_2^2. \tag{4}$$

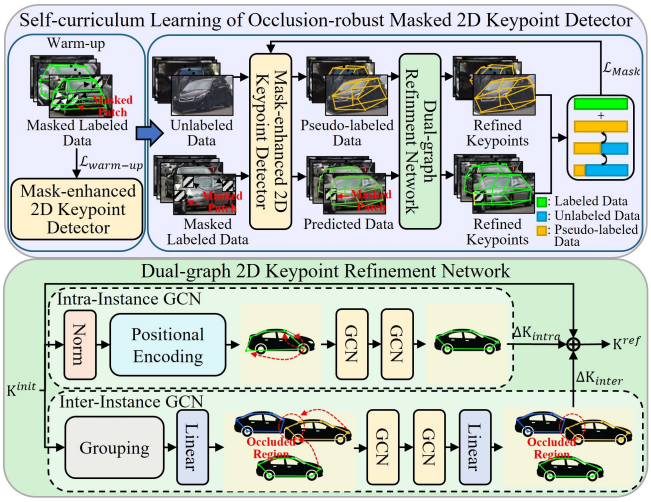

Figure 4: Occlusion-robust masked 2D skeleton keypoint detector

Trained with the LoRA loss, the Stable Diffusion model performs L purification iterations. Ultimately, through this process, we obtain z_0^p.

Segmentation Step. To support 3D object detection, the proposed module predicts a segmentation feature and employs it as a guide. The segmentation branch is derived from the encoder output, normalized using a convolutional network with group normalization, and passed through the segmentation decoder to obtain f_s. We then multiply f_s by the target-region probability map R to emphasize detection-critical regions. Segmentation can occasionally be unreliable under strong attacks; thus, this work introduces a confidence-driven filtering mechanism based on the SE block. This approach inherently derives purified features through disentangled representations and secures a reliable segmentation feature by utilizing an SE block to preemptively suppress erroneous extractions during 3D localization.

3.3 3D Detection Decoder with Occlusion-robust 2D Skeleton Keypoint

This section explains how MonoPure predicts and refines occlusion-robust 2D skeleton keypoints for 3D localization. This module is built around three key elements, namely a mask-enhanced 2D keypoint detector, self-curriculum learning, and dual-graph 2D keypoint refinement. MonoPure separates the 2D and 3D detection decoders to balance reliable 2D localization and geometry-aware 3D reasoning, using f_s as a shared structural cue. The visual encoder extracts appearance features for 2D localization, whereas the depth encoder provides depth features for geometry-aware reasoning. The 2D decoder further predicts occlusion-robust 2D skeleton keypoints and object-level spatial cues. The refined 2D keypoints are then injected into the 3D decoder through depth-KP cross-attention, serving as object-level geometric anchors for 3D localization.

Mask-enhanced 2D Skeleton Keypoint Detector. This work adopts CenterNet [Duan *et al.*, 2019] as the 2D detection decoder and keypoint heatmap-based detector. Using the 2D bounding boxes predicted by the 2D detection decoder as priors, the detector predicts I keypoint heatmaps $\{\hat{H}_i\}_{i=1}^{I}$ with $\hat{H}_i \in \mathrm{R}^{S\times S}$. This approach initializes the mask-enhanced 2D keypoint detector on labeled source data $\mathcal{D}_L$ by minimizing the mean squared error (MSE) between predicted heatmaps and Gaussian heatmaps rendered from ground-truth keypoints H. Next, this method improves occlusion robustness by initializing training with mask augmentation, sampling a rectangular occlusion mask $m \sim \mathcal{P}_{occ}$, where $\mathcal{P}_{occ}$ reflects source-domain occlusion statistics, and applying $\mathcal{A}(x;m)$ to zero out the masked region. The initialization objective is:

$$\mathcal{L}_{\text{warm-up}} = \mathbb{E}_{H\sim\mathcal{D}_L}\left[\mathbb{E}_{m\sim\mathcal{P}_{occ}}\left[\frac{1}{K}\left\|\hat{H}(\mathcal{A}(x;m)) - H\right\|_2^2\right]\right]. \tag{5}$$

Following this initialization, a curriculum learning strategy is adopted that progressively adapts the model from labeled source data to more challenging unlabeled target scenarios. This scheme applies the detector to the unlabeled target set $\mathcal{D}_U$ and refines the raw predictions using the dual-graph refinement network (DGR-Net). This approach retains pseudo-labeled keypoints as $S(x) = \{i \mid \sigma_i > 0.5\}$ with $K_x = |S(x)|$ and retrains the detector using both labeled and pseudo-labeled data:

$$\mathcal{L}_{\text{mask}} = \mathcal{L}_{\text{warm-up}} + \mathcal{L}_{\text{pseudo}}. \tag{6}$$

Keypoint pseudo-labels are filtered by confidence. For unlabeled data, masking was disabled during self-training to avoid amplifying pseudo-label noise. The pseudo-label loss is:

$$\mathcal{L}_{\text{pseudo}} = \mathrm{E}_{x\sim\mathcal{D}_U}\left[\frac{1}{K_x}\sum_{k\in\mathcal{S}(x)}\left\|\hat{H}_k(x) - H_k^{\text{pseudo}}\right\|_2^2\right]. \tag{7}$$

This iterative curriculum learning process ensures stabilization on reliable pseudo-labels and continues until the entire dataset is labeled. All training is done off-line, and during inference, this approach applies only the mask-enhanced 2D keypoint detector and DGR-Net using trained models.

Dual-Graph 2D Keypoint Refinement. DGR-Net refines initial 2D keypoints via a hierarchical spatial reasoning strategy that models both local geometric constraints and global scene context. This dual-branch architecture is motivated by the fact that keypoint accuracy depends not only on an object's rigid topology but also on its spatial interactions with neighboring instances, especially under occlusion.

Given the initial keypoints $K^{\text{init}} = \{\hat{k}_i^{\text{init}}\}_{i=1}^{K}$, this approach builds an object-specific graph whose nodes are keypoints and whose edges follow a fixed adjacency matrix defined by the canonical skeletal topology. Using this structural prior, a lightweight GCN propagates geometric cues to predict intra-instance offsets $\Delta K_{\text{intra}} = \{\Delta\hat{k}_{\text{intra},i}\}_{i=1}^{K}$.

This approach builds a dynamic neighbor graph for each image to resolve ambiguities induced by scene-level interactions. Each node represents an object instance, and the edges are adaptively established between spatially proximal objects based on 2D centroid distances. Features from these

neighboring nodes are aggregated via a GCN to compute inter-instance offsets: $\Delta K_{\text{inter}} = \{\Delta \hat{k}_{\text{inter},i}\}_{i=1}^{K}$.

The final refined keypoints are computed as $K^{\text{ref}} = K^{\text{init}} + \Delta K_{\text{intra}} + \Delta K_{\text{inter}}$. This work follows the standard GCN [Kipf and Welling, 2017] protocols regarding layer depth and neighborhood definitions to maintain lightweight computation while ensuring training convergence.

DGR-Net is trained exclusively on labeled data $\mathcal{D}_L$ by minimizing the Smooth ℓ_1 loss between the refined keypoints and the ground-truth keypoints K^{gt}. We adopt this loss function to strike a balance between robustness against large coordinate outliers and stable, oscillation-free convergence for small residual errors during the fine-grained refinement process, which is formulated as follows:

$$\mathcal{L}_{\text{refine}} = \mathbb{E}_{(x,K^{\text{gt}})\sim\mathcal{D}_L}\left[\frac{1}{K}\text{Smooth-}\ell_1\left(K^{\text{ref}} - K^{\text{gt}}\right)\right]. \quad (8)$$

3D Detection Decoder. The 3D detection decoder integrates depth cues and structural priors via depth and depth-KP cross-attention layers. The depth-KP cross-attention layer applies the refined keypoints K^{ref} from the previous stage. This approach employs the depth features $f_d \in \mathbb{R}^{N\times D}$ as queries and refined keypoints $K^{\text{ref}} \in \mathbb{R}^{N\times D}$ as keys or values, where N denotes the number of feature tokens. The cross-attention yields the refined depth feature $\tilde{f}_d$ as follows:

$$\tilde{f}_d = \text{softmax}\left(\frac{(f_d \cdot W_Q)\cdot\left(K^{\text{ref}}\cdot W_K\right)^{\top}}{\sqrt{D}}\right)\cdot\left(K^{\text{ref}}\cdot W_V\right), \quad (9)$$

where D denotes the feature (embedding) dimension. Moreover, W_Q, W_K and $W_V \in \mathbb{R}^{D\times D}$ are learnable linear projection matrices for the query, key, and value, respectively.

Next, the 3D detection decoder $\text{Dec}(\cdot)$, which consists of inter-instance self-attention, visual cross-attention, and a feed-forward network, is applied to compute the final 3D box X_{3D} as follows:

$$X_{3D} = \text{Dec}\left(f_v, \tilde{f}_d, \phi\right), \quad (10)$$

where $\phi \in \mathbb{R}^{3\times 4}$ denotes the camera calibration parameters, f_v denotes the visual encoder feature.

3.4 Projective Calib-purification

Although the 3D decoder enforces internal geometric consistency, accurate calibration is critical for 3D localization. In monocular detection, calibration bridges the 3D scene and 2D image plane; thus, adversarial distortions can cause severe localization errors even when image features are extracted well [Veicht *et al.*, 2024]. Projecting 3D predictions onto the 2D plane visualizes mismatches with reliable image-based anchors. Therefore, purifying corrupted calibration parameters is crucial to restoring geometric alignment for robust 3D localization.

The proposed method relies on the fact that image-based 2D detection results remain invariant to changes in camera calibration parameters. For each 3D vertex with homogeneous coordinates $X_{3D} = [X, Y, Z, 1]^T$, the 2D pixel coordinate $X_{2D} = [u, v]^T$ is obtained by the perspective projection:

$$X_{2D} = \pi(\phi \cdot X_{3D}), \quad \phi = \begin{bmatrix} f_x & 0 & c_x & 0 \\ 0 & f_y & c_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad (11)$$

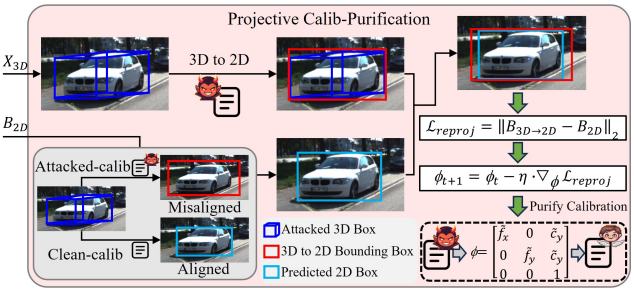

Figure 5: Illustration of projective calib-purification

where $\pi(\cdot)$ denotes the perspective projection operator that performs depth scaling to obtain the final pixel coordinates.

The eight projected vertices are determined using X_{2D} and are employed to derive the 2D axis-aligned bounding box $B_{3D\to 2D}$, representing the 2D footprint of the predicted 3D structure. The coordinates of B_{3D} are transformed into eight corner coordinates, which are projected onto 2D coordinates using Eqs. (12) and (13), yielding $\{u_j\}_{j=1}^{8}$ and $\{v_i\}_{i=1}^{8}$. Based on these projected values, the projected 2D bounding box $B_{3D\to 2D}$ is defined to encompass the corners of B_{3D} as follows:

$$B_{3D\to 2D} = (u_j, v_i, w, h), \quad (12)$$

$$u_j = \frac{u_{\max} + u_{\min}}{2}, \quad v_i = \frac{v_{\max} + v_{\min}}{2}, \quad w = u_{\max} - u_{\min}, \quad h = v_{\max} - v_{\min}, \quad (13)$$

where $u_{\min}$ and $u_{\max}$ denote the minimum and maximum u-values of u_j. Further, $v_{\min}$ and $v_{\max}$ denote the minimum and maximum v-values of v_i. This method minimizes reprojection loss by enforcing intrinsic constraints that strictly restrict all camera parameters from exceeding the physical boundaries of the image resolution. It ensures focal lengths and the principal point stay within the image frame size. To enhance the consistency between $B_{3D\to 2D}$ and B_{2D} and reduce outlier corners outside B_{2D}, the loss is expressed as follows:

$$\mathcal{L}_{reproj} = \|B_{3D\to 2D} - B_{2D}\|_2. \quad (14)$$

Crucially, errors in 2D detections do not affect the 3D-to-2D projection, enhancing robustness. This work updates the parameters until the reprojection loss converges using backpropagation $\mathcal{L}_{reproj}$ via the differentiable projection layer:

$$\phi_{t+1} = \phi_t - \eta \cdot \nabla_\phi \mathcal{L}_{reproj}, \quad (15)$$

where $\eta = 10^{-4}$ represents the learning rate, and ϕ_{t+1} denotes the purified parameters for the next step. The iterative process terminates after the reprojection loss has sufficiently converged. Despite severe system-level distortions, MonoPure purifies calibration and ensures robust 3D localization by minimizing the 3D-to-2D projection gap.

4 Experiments

4.1 Dataset

The experiments were conducted on two autonomous-driving datasets: KITTI [Geiger *et al.*, 2013] and nuScenes [Caesar *et al.*, 2020]. The KITTI dataset contains 7,481 annotated

$AP_{3D\|R40}(\%)\uparrow$															
Methods	Single-Component Adv. (Image)						Multi-Component Adv. (Image and Calib.)						Clean		
	MI-FGSM			PGD			MI-FGSM			PGD			No Attack		
	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
*DEVIANT [Kumar *et al.*, 2022]	10.92	6.92	5.68	13.62	8.79	6.66	< 0.01 (7.89)	< 0.01 (5.05)	< 0.01 (3.50)	0.01 (10.30)	< 0.01 (6.42)	< 0.01 (4.81)	24.73	16.58	14.53
*MonoDTR [Huang *et al.*, 2022]	14.76	10.39	8.06	17.89	12.70	10.03	< 0.01 (11.37)	< 0.01 (7.33)	< 0.01 (6.02)	< 0.01 (14.81)	< 0.01 (9.56)	< 0.01 (7.74)	24.57	18.60	15.29
*MonoDETR [Zhang *et al.*, 2023]	19.23	12.50	10.36	19.82	13.29	10.89	< 0.01 (14.08)	< 0.01 (9.63)	< 0.01 (7.32)	0.01 (16.49)	0.01 (11.43)	0.01 (8.97)	28.09	20.82	17.46
*MonoCD [Yan *et al.*, 2024]	17.44	11.56	9.38	19.19	12.74	11.02	0.01 (13.63)	< 0.01 (7.86)	< 0.01 (7.04)	< 0.01 (15.28)	< 0.01 (10.82)	< 0.01 (8.36)	26.45	19.37	16.38
*MonoDGP [Pu *et al.*, 2025]	19.98	13.98	11.27	22.23	15.95	13.21	< 0.01 (16.19)	< 0.01 (11.83)	< 0.01 (9.91)	< 0.01 (18.04)	0.01 (12.81)	0.01 (10.83)	30.12	22.68	19.38
Ours	**24.15**	**17.47**	**14.53**	**26.09**	**18.94**	**17.31**	**20.48**	**14.97**	**12.16**	**22.76**	**16.67**	**15.86**	**31.29**	**23.00**	**19.78**

$AP_{BEV\|R40}(\%)\uparrow$															
Methods	Single-Component Adv. (Image)						Multi-Component Adv. (Image and Calib.)						Clean		
	MI-FGSM			PGD			MI-FGSM			PGD			No Attack		
	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
*DEVIANT [Kumar *et al.*, 2022]	15.92	10.72	8.42	20.08	13.40	10.68	0.20 (14.59)	0.19 (9.52)	0.10 (8.04)	0.26 (19.36)	0.22 (12.88)	0.10 (10.30)	32.60	23.05	19.99
*MonoDTR [Huang *et al.*, 2022]	22.21	15.81	12.22	25.14	14.63	12.33	0.74 (19.57)	0.74 (13.49)	0.70 (9.81)	0.54 (23.29)	0.57 (16.93)	0.45 (12.65)	34.11	25.19	21.30
*MonoDETR [Zhang *et al.*, 2023]	26.39	18.13	14.77	28.33	19.40	15.89	0.55 (24.64)	0.42 (16.56)	0.39 (13.28)	0.60 (27.26)	0.48 (18.88)	0.48 (15.40)	37.59	27.28	23.37
*MonoCD [Yan *et al.*, 2024]	25.71	16.95	13.70	26.19	18.16	13.92	0.23 (23.49)	0.29 (16.09)	0.17 (12.62)	0.59 (26.17)	0.22 (18.78)	0.19 (14.21)	34.60	24.96	21.51
*MonoDGP [Pu *et al.*, 2025]	27.67	19.15	15.87	29.78	21.12	17.83	0.25 (26.42)	0.29 (19.03)	0.27 (15.63)	0.20 (28.82)	0.23 (20.47)	0.20 (17.17)	39.88	29.13	25.21
Ours	**30.83**	**22.46**	**19.86**	**34.20**	**25.23**	**21.09**	**30.45**	**22.38**	**19.02**	**30.51**	**22.43**	**19.47**	**40.37**	**29.33**	**25.67**

Table 1: Results of monocular 3D object detectors on the KITTI car class under two adversarial attacks (MI-FGSM and PGD) across three evaluation settings (single-component adversarial, multi-component adversarial, and clean) at an IoU threshold of 0.7.

images, with 3,712 images for training and 3,769 images for validation, following the standard split. The nuScenes dataset was employed under a monocular setting with front-view camera images, where 28,130 images were for training and 6,019 images for validation. Following DEVIANT [Kumar *et al.*, 2022], this work adopts a KITTI-style data split for nuScenes to enable a fair evaluation protocol. Across all datasets, only monocular RGB images and camera calibration information were used during training and inference. The experimental results were evaluated using the 3D average precision metric with 40 recall positions.

4.2 Experiment Settings

Training Settings. MonoPure was trained for 95 epochs with an initial learning rate of 2e-4 and a batch size of 1 on a single RTX-4080 GPU, using the AdamW optimizer with weight decay. In the disentangled image-purification and segmentation module, the disentangling encoder adopted ResNet50 [He *et al.*, 2016] as the feature backbone. For diffusion model training, Stable Diffusion [Rombach *et al.*, 2022] was employed as the base model and fine-tuned via LoRA [Hu *et al.*, 2022] with the TCD scheduler [Zheng *et al.*, 2024]. Training was conducted with a batch size of 2 for 1,000 steps, using an initial learning rate of 1e-4. During this process, only the LoRA parameters were updated. With the LoRA fine-tuned Stable Diffusion model, the number of purification iterations is set to L=1, and additional analyses under varying L are provided in *Appendix A*.

Attack Settings. This work considers a strong adversarial setting in which the attacker has full access to the input images, camera calibration parameters, visual/depth encoder, and 2D/3D detection decoder. This work uses MI-FGSM [Dong *et al.*, 2018] and PGD [Madry *et al.*, 2017], widely used adversarial attacks, to attack the input images and camera calibration parameters for adversarial evaluation. The perturbation strength is controlled by an ℓ_∞ norm bound with $\epsilon = 32$, following standard practice in adversarial attacks. Unlike an adversarial learning approach, MonoPure performs purification without any prior knowledge of the attack, and its purification remains effective against unseen adversarial attacks. *Appendix A* provides additional comparison results on universal adversarial perturbations (UAP) attacks [Moosavi-Dezfooli *et al.*, 2017], as well as results for a wider range of ϵ values.

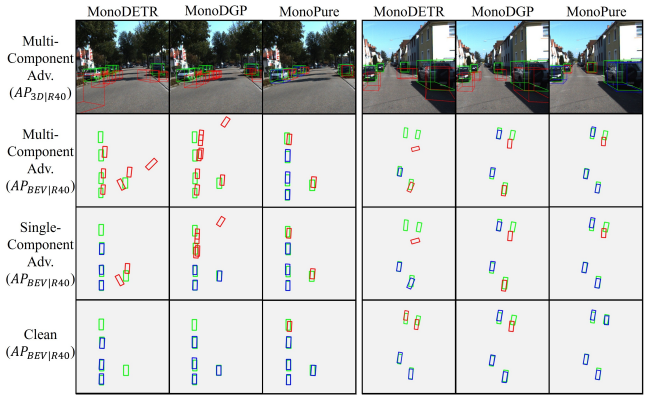

Figure 6: Visual evaluation of the monocular 3D object detectors, MonoDGP, MonoDETR and MonoPure, on KITTI at on IoU threshold of 0.5 under MI-FGSM attacks. True detection: blue; false detection: red; ground truth: green.

4.3 Result and Analysis

All experiments apply image purification under image attacks, calibration purification under calibration attacks, and no purification under clean settings. For fair comparison, this work reports results of prior monocular 3D detectors on purified images. Methods marked with * use Stable Diffusion with the TCD scheduler for purification when image attacks are present. For multi-component attacks, we additionally report prior detectors with MonoPure's calib-purification denoted in parentheses $(\cdot)$, as no existing method purifies attacked camera calibration parameters. Bold text indicates the best, and underlined text indicates the second-best results.

Table 1 reveals that the proposed method achieved the best accuracy across adversarial and clean settings. The approach surpasses prior SOTA across three difficulty levels. Under a single-component attack, across attack methods, MonoPure improved over the second-best method by an average of $3.68\%p$ across difficulty levels on AP_{3D} and AP_{BEV}. Notably, the multi-component attack is highly destructive. All prior methods collapsed to near-zero performance. In contrast, the proposed approach retained strong detection per-

$AP_{3D\|R40}(\%)\uparrow$					
Methods	Single-Component Adv.		Multi-Component Adv.		Clean
	MI-FGSM	PGD	MI-FGSM	PGD	No Attack
*DEVIANT	12.39	15.87	0.10 (9.48)	0.12 (12.60)	26.49
*MonoDTR	15.45	19.38	0.13 (12.29)	0.15 (16.19)	27.09
*MonoDETR	18.02	21.73	0.17 (15.94)	0.18 (17.88)	28.72
*MonoCD	17.36	21.79	0.11 (15.36)	0.16 (17.62)	27.76
*MonoDGP	19.15	23.15	0.19 (16.32)	0.21 (19.76)	28.78
Ours	**22.02**	**25.33**	**18.67**	**23.16**	**29.39**
$AP_{BEV\|R40}(\%)\uparrow$					
Methods	Single-Component Adv.		Multi-Component Adv.		Clean
	MI-FGSM	PGD	MI-FGSM	PGD	No Attack
*DEVIANT	13.11	15.17	0.21 (11.15)	0.33 (13.54)	27.87
*MonoDTR	17.56	17.88	0.41 (14.81)	0.44 (16.23)	28.48
*MonoDETR	21.01	24.74	0.39 (18.62)	0.42 (22.09)	29.55
*MonoCD	18.12	22.89	0.30 (16.42)	0.33 (19.89)	29.32
*MonoDGP	21.59	25.94	0.43 (21.52)	0.47 (23.41)	31.64
Ours	**24.10**	**27.87**	**23.89**	**26.80**	**32.13**

Table 2: Results of monocular 3D object detectors on the nuScenes for the car class under two adversarial attacks (MI-FGSM and PGD), evaluated at an IoU threshold of 0.5 on the Moderate setting.

formance, with pronounced gains in AP_{BEV}. The method consistently maintains strong performance under adversarial attacks and clean settings. *Appendix A* provides additional quantitative results on pedestrian class, showing that MonoPure achieves SOTA performance with the same trend as in Table 1.

In addition to the quantitative analysis, Figure 6 presents qualitative visual comparisons among MonoDETR, MonoDGP and MonoPure under adversarial and clean settings. Under the multi-component adversarial attack, where the camera calibration parameter is adversarially perturbed, MonoDETR and MonoDGP displayed pronounced geometric distortions, with the 3D bounding boxes in the top KITTI image appearing vertically misaligned. Likewise, under the single-component adversarial image attack, the BEV visualizations revealed that image-space perturbations substantially disrupt object detection in MonoDETR and MonoDGP, leading to missing 3D bounding boxes or inaccurate localization. In contrast, MonoPure remained robust to calibration and image attacks, producing stable predictions and 3D bounding boxes that remained closely aligned with the ground truth annotations. Moreover, under the clean setting, MonoPure consistently demonstrates strong performance, remaining particularly robust in occluded scenarios by using keypoint-based spatial cues. *Appendix B* provides additional qualitative visualizations of detection results and skeleton keypoints on the KITTI and nuScenes datasets.

To verify that the observed robustness trends are consistent beyond KITTI, this work evaluates MonoPure on nuScenes (Table 2). Consistent with the results in Table 1, MonoPure achieved SOTA performance across all attack scenarios and in clean settings. Notably, under the multi-component attack, MonoPure retained meaningful detection accuracy, yielding dramatic performance improvement. Limitations and future work are discussed in *Appendix D*.

4.4 Efficiency

Figure 7 compares the complexity of monocular 3D object detectors, evaluated in terms of the number of training parameters, FLOPs, and latency. MonoPure maintained a comparable number of training parameters to prior methods due

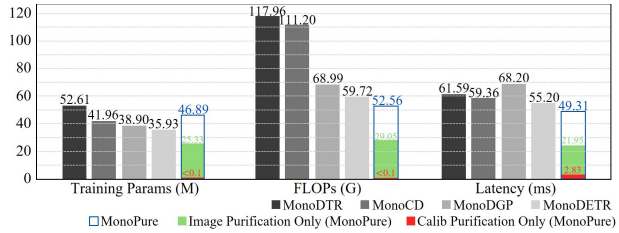

Figure 7: Computational complexity comparison on the KITTI.

Disentangled Image-purification and Segmentation	3D Detection Decoder using Occlusion-robust Keypoint	Projective Calib-purification	Multi-component Adv. $AP_{3D\|R40}(\%)\uparrow$ Easy	Mod.	Hard	$AP_{BEV\|R40}(\%)\uparrow$ Easy	Mod.	Hard
✓	✓	✓	20.48	14.97	12.16	30.45	22.38	19.02
	✓	✓	12.41	8.64	7.20	15.09	12.74	10.63
✓		✓	18.72	11.69	11.85	28.37	21.97	18.91
✓	✓		0.01	< 0.01	< 0.01	0.80	0.65	0.27
			< 0.01	< 0.01	< 0.01	0.25	0.29	0.27

Table 3: Ablation study for MonoPure on the KITTI, evaluated at an IoU threshold of 0.7 under MI-FGSM attacks.

to LoRA. Although MonoPure uses slightly more parameters, its model size remains practical for deployment on in-vehicle automotive platforms (e.g., NVIDIA DRIVE AGX Orin), enabling real-time inference with around 50M parameters. Notably, MonoPure achieved the lowest FLOPs and latency, making it suitable for real-time deployment. This efficiency is fostered by performing purification and detection at half-resolution of input images, parallelizing the visual encoder and depth encoder, and sharing a 2D detection decoder with the 2D keypoint detector.

4.5 Ablation Study

Each component is crucial under the multi-component adversarial attack (Table 3). Removing the disentangled image-purification and segmentation module or the projective calibration purification caused a catastrophic decrease in AP_{3D}, to under 1, indicating that image and calibration purification are necessary for reliable 3D detection. In addition, the occlusion-robust keypoint-based 3D detection decoder consistently improved performance, confirming its role in mitigating occlusion-induced ambiguities. *Appendix A* provides additional ablation studies on the nuScenes dataset, which show trends similar to those observed on KITTI. It also reports detailed quantitative evaluations for each of the three components ablated in Table 3.

5 Conclusion

This paper addresses the vulnerability of monocular 3D object detection to challenging multi-component adversarial attacks, where images and camera calibration parameters are compromised. To address this problem, this work proposes MonoPure, which performs detection-aligned image purification by combining a disentanglement-based design with a target-region probability map. Moreover, MonoPure presents a 3D detection decoder that improves structural robustness under occlusion by using 2D skeleton keypoints. Finally, it purifies corrupted camera calibration parameters via a 2D box reprojection constraint. Across diverse attack settings, including multi-component adversarial attacks, MonoPure achieved real-time SOTA performance.

Acknowledgements

This work was supported by the IITP grant funded by the Korea government (MSIT) (RS-2022-00156287, RS-2023-00256629, RS-2024-00437718, RS-2026-25619158).

Contribution Statement

Yeon Woo Cho, Jung Woo Cheon, Seung-hyeok Back contributed equally to this work.

References

[Caesar *et al.*, 2020] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[Chen *et al.*, 2016] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun. Monocular 3d object detection for autonomous driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2147–2156, 2016.

[Chen *et al.*, 2024] Hai Chen, Huanqian Yan, Xiao Yang, Hang Su, Shu Zhao, and Fulan Qian. Efficient adversarial attack strategy against 3d object detection in autonomous driving systems. *IEEE Transactions on Intelligent Transportation Systems*, 25(11):16118–16132, 2024.

[Cheng *et al.*, 2022] Zhiyuan Cheng, James Liang, Hongjun Choi, Guanhong Tao, Zhiwen Cao, Dongfang Liu, and Xiangyu Zhang. Physical attack on monocular depth estimation with optimal adversarial patches. In *European conference on computer vision*, pages 514–532. Springer, 2022.

[Choi *et al.*, 2024] Daewon Choi, Jongheon Jeong, Huiwon Jang, and Jinwoo Shin. Adversarial robustification via text-to-image diffusion models. In *European Conference on Computer Vision*, pages 158–177. Springer, 2024.

[Dong *et al.*, 2018] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018.

[Duan *et al.*, 2019] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6569–6578, 2019.

[Gambashidze *et al.*, 2024] Alexander Gambashidze, Aleksandr Dadukin, Maxim Golyadkin, Maria Razzhivina, and Ilya Makarov. Do you remember... the future? weak-to-strong generalization in 3d object detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 8653–8656, 2024.

[Geiger *et al.*, 2013] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11):1231–1237, 2013.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Hinton and Salakhutdinov, 2006] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[Hu *et al.*, 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[Hu *et al.*, 2022] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[Huang *et al.*, 2022] 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*, pages 4012–4021, 2022.

[Kang *et al.*, 2024] Caixin Kang, Yinpeng Dong, Zhengyi Wang, Shouwei Ruan, Yubo Chen, Hang Su, and Xingxing Wei. Diffender: Diffusion-based adversarial defense against patch attacks. In *European Conference on Computer Vision*, pages 130–147. Springer, 2024.

[Kipf and Welling, 2017] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[Kou *et al.*, 2023] Ziyi Kou, Shichao Pei, Yijun Tian, and Xiangliang Zhan. Character as pixels: A controllable prompt adversarial attacking framework for black-box text guided image generation models. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)*. Proceedings of the Thirty-Second International Joint Conference on . . . , 2023.

[Kumar *et al.*, 2022] Abhinav Kumar, Garrick Brazil, Enrique Corona, Armin Parchami, and Xiaoming Liu. Deviant: Depth equivariant network for monocular 3d object detection. In *ECCV*, pages 664–683. Springer, 2022.

[Lee *et al.*, 2024] Eun-Gi Lee, Moon Seok Lee, Jae Hyun Yoon, and Seok Bong Yoo. Intenspure: attack intensity-aware secondary domain adaptive diffusion for adversarial purification. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, Jeju, Republic of Korea*, pages 3–9, 2024.

[Lei *et al.*, 2025] Chun Tong Lei, Hon Ming Yam, Zhongliang Guo, Yifei Qian, and Chun Pong Lau. Instant adversarial purification with adversarial consistency distillation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24331–24340, 2025.

[Li *et al.*, 2023] Xingyuan Li, Jinyuan Liu, Long Ma, Xin Fan, and Risheng Liu. Advmono3d: Advanced monocu-

lar 3d object detection with depth-aware robust adversarial training. *arXiv preprint arXiv:2309.01106*, 2023.

[Lin *et al.*, 2024] Hongbin Lin, Yifan Zhang, Shuaicheng Niu, Shuguang Cui, and Zhen Li. Monotta: Fully test-time adaptation for monocular 3d object detection. In *European Conference on Computer Vision*, pages 96–114. Springer, 2024.

[Liu and Jiang, 2025] Yimin Liu and Peng Jiang. Generic adversarial attack framework against vertical federated learning. In *34th International Joint Conference on Artificial Intelligence, IJCAI 2025*, pages 5806–5814. International Joint Conferences on Artificial Intelligence, 2025.

[Liu *et al.*, 2025] Hou-I Liu, Christine Wu, Jen-Hao Cheng, Wenhao Chai, Shian-Yun Wang, Gaowen Liu, Hugo Latapie, Jhih-Ciang Wu, Jenq-Neng Hwang, Hong-Han Shuai, et al. Monotakd: Teaching assistant knowledge distillation for monocular 3d object detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22266–22275, 2025.

[Madry *et al.*, 2017] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[Moosavi-Dezfooli *et al.*, 2017] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.

[Niu and Yang, 2023] Zhong-Han Niu and Yu-Bin Yang. Defense against adversarial attacks with efficient frequency-adaptive compression and reconstruction. *Pattern Recognition*, 138:109382, 2023.

[Pu *et al.*, 2025] Fanqi Pu, Yifan Wang, Jiru Deng, and Wenming Yang. Monodgp: Monocular 3d object detection with decoupled-query and geometry-error priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6520–6530, 2025.

[Qi *et al.*, 2025] Yafei Qi, Menghao Yang, Yongmin Zhang, Bing Xiong, Yaping Liu, and Zhaoning Zhang. Selfdn: Adaptive self-denoising for multi-view 3d object detection. *Knowledge-Based Systems*, page 113816, 2025.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[Shi *et al.*, 2025] Peicheng Shi, Xinlong Dong, Runshuai Ge, Zhiqiang Liu, and Aixi Yang. Dp-m3d: Monocular 3d object detection algorithm with depth perception capability. *Knowledge-Based Systems*, 318:113539, 2025.

[Veicht *et al.*, 2024] Alexander Veicht, Paul-Edouard Sarlin, Philipp Lindenberger, and Marc Pollefeys. Geocalib: Learning single-image calibration with geometric optimization. In *European Conference on Computer Vision*, pages 1–20. Springer, 2024.

[Xue *et al.*, 2025] Ziteng Xue, Mingzhe Guo, Heng Fan, Shihui Zhang, and Zhipeng Zhang. Corrbev: Multi-view 3d object detection by correlation learning with multimodal prototypes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27413–27423, 2025.

[Yan *et al.*, 2024] Longfei Yan, Pei Yan, Shengzhou Xiong, Xuanyu Xiang, and Yihua Tan. Monocd: Monocular 3d object detection with complementary depths. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10248–10257, 2024.

[Yang *et al.*, 2021] Shuo Yang, Tianyu Guo, Yunhe Wang, and Chang Xu. Adversarial robustness through disentangled representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3145–3153, 2021.

[Yoon *et al.*, 2021] Jongmin Yoon, Sung Ju Hwang, and Juho Lee. Adversarial purification with score-based generative models. In *International Conference on Machine Learning*, pages 12062–12072. PMLR, 2021.

[Zhang *et al.*, 2023] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9155–9166, 2023.

[Zhang *et al.*, 2025] Xiayue Zhang, Huashuo Lei, Daizong Liu, Xiaoye Qu, Xiang Fang, Runwei Guan, and Keyan Jin. Monoattack: A strong attack framework with depth-migration and attribute-tampering for monocular 3d object detection. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2025.

[Zheng *et al.*, 2024] Jianbin Zheng, Minghui Hu, Zhongyi Fan, Chaoyue Wang, Changxing Ding, Dacheng Tao, and Tat-Jen Cham. Trajectory consistency distillation: Improved latent consistency distillation by semi-linear consistency function with trajectory mapping, 2024.

[Zhou *et al.*, 2021] Yunsong Zhou, Yuan He, Hongzi Zhu, Cheng Wang, Hongyang Li, and Qinhong Jiang. Monocular 3d object detection: An extrinsic parameter free approach. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7556–7566, 2021.

[Zhu *et al.*, 2023] Zijian Zhu, Yichi Zhang, Hai Chen, Yinpeng Dong, Shu Zhao, Wenbo Ding, Jiachen Zhong, and Shibao Zheng. Understanding the robustness of 3d object detection with bird's-eye-view representations in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21600–21610, 2023.