

Article

Lightweight Monocular Distance Estimation via Anisotropic Geometry Loss for Low-Light Driving Environments

Ricky Christanto  and Shaou-Gang Miaou * 

Department of Electronic Engineering, Chung Yuan Christian University, Taoyuan 320314, Taiwan;
g11202603@cycu.edu.tw

* Correspondence: miaou@cycu.edu.tw

Highlights

What are the main findings?

- We introduce the Anisotropic Geometry Loss (AGL) framework, a 2.71 M-parameter monocular distance estimator that achieves >160 FPS while reducing Dark-KITTI RMSE from 11.53 m (Vanilla YOLOv10n) to 10.55 m with the optional CLAHE preprocessor—without modifying the YOLOv10n backbone.
- Joint sensitivity analysis of the loss coefficients (λ_y, λ_h) reveals that independent single-axis tuning misidentifies the global optimum and exposes seed-instability at apparent minima—motivating cross-seed-stable coefficient selection over greedy single-axis tuning for lightweight monocular 3D detectors.

What are the implications of the main findings?

- Explicit pinhole-geometry constraints, rather than heavyweight architectural depth-prediction heads, are sufficient to suppress the multi-meter monocular depth errors caused by floating bounding boxes, enabling camera-only Autonomous Emergency Braking (AEB) to demonstrate computational efficiency on desktop GPU and provide a promising foundation for future embedded deployment aligned with Euro NCAP safety protocols.
- The Far-zone (>30 m) results demonstrate that all evaluated methods converge to a similar localization accuracy, reflecting the fundamental limitation of static monocular projection rather than model-specific performance. The proposed AGL framework therefore delivers its principal benefit in the Medium zone (15–30 m), while maintaining equivalent Far-zone performance to existing techniques. This behavior is consistent with Euro NCAP monitoring-only requirements beyond 30 m and suggests that future improvements for long-range localization should incorporate dynamic camera-pose estimation and graded-terrain compensation.



Academic Editor: Qiong Wu

Received: 1 June 2026

Revised: 29 June 2026

Accepted: 10 July 2026

Published: 13 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Abstract

Robust monocular distance estimation under varying illumination conditions is critical for autonomous driving safety. While state-of-the-art monocular 3D detection models achieve high accuracy in daylight conditions, they rely on computationally heavy architectures and degrade significantly in low-light environments. Lightweight 2D detectors (e.g., YOLO variants) offer real-time performance but lack the geometric constraints required for accurate depth estimation. To address this limitation, we propose the Anisotropic Geometry Loss (AGL) framework. This lightweight framework enforces ground-plane consistency through an anisotropic bottom-edge constraint derived from the pinhole camera model. In addition, a luminance-channel contrast enhancement module (CLAHE) is

applied at inference to improve low-light visibility. Experimental results on the Dark-KITTI dataset show that the proposed method achieves an RMSE of 10.91 ± 0.68 m, improving over YOLOv10n (11.53 ± 0.56 m) and YOLOv26n (11.99 ± 0.58 m), while maintaining a 2.71 M-parameter footprint and real-time inference (>160 FPS). With CLAHE, RMSE is further reduced to 10.55 ± 0.72 m. Stratified by kinematic safety zone, the proposed method achieves 2.42 ± 0.03 m in the Near range (0–15 m), 5.94 ± 0.19 m in the Medium range (15–30 m), and 17.41 ± 1.25 m in the Far range (>30 m), corresponding to Euro NCAP AEB (Autonomous Emergency Braking) stopping distances. AGL provides its largest measurable accuracy improvement in the medium-distance range while maintaining comparable performance in the far-distance range. A complementary luminance-channel CLAHE pre-processor recovers bottom-edge gradients in synthetic and real low-light frames; zero-shot generalization is qualitatively corroborated on the ExDark dataset. These results demonstrate that explicit geometric constraints provide an effective and efficient solution for robust cross-illumination resistance in monocular distance estimation. The framework also shows practical potential for camera-only AEB systems deployed on edge-computing platforms aligned with Euro NCAP safety protocols.

Keywords: monocular distance estimation; illumination-invariant vision; autonomous emergency braking; efficient GPU computing; nocturnal autonomous driving

1. Introduction

The transition of autonomous vehicles (AVs) from highly controlled operational design domains to unstructured, real-world environments requires perceptual systems that remain robust under severe visual degradation [1]. Nighttime driving remains disproportionately dangerous, as standard camera sensors suffer from low signal-to-noise ratios, motion blur, and headlight glare [2]. In driving environments, illumination degradation and exposure imbalance can significantly affect scene perception and safety-related decision making. Previous studies have investigated exposure correction techniques specifically for driving scenarios to improve visual quality under challenging lighting conditions [3].

Our work targets lightweight computational efficiency on desktop GPU and provides a promising foundation for future embedded deployment. A sub-3 M-parameter detector operating under real-time AEB latency budgets—where heavyweight monocular 3D architectures are not viable design candidates.

As illustrated in Figure 1, transitioning from diurnal to nocturnal environments significantly degrades high-frequency spatial features. In the daytime baseline (top), the tire-road interface is clearly delineated, allowing for stable geometric anchoring. However, in the nocturnal domain (bottom), low-light noise and reduced dynamic range cause the vehicle's lower boundary to submerge into the road's luminance floor, resulting in the so-called 'floating bounding box' phenomenon.

Monocular distance estimation relies on accurate extraction of the tire-road interface for geometric anchoring. As shown in the transition from KITTI (day) to the Dark-KITTI (night) variant, low-light conditions and an elevated sensor noise floor result in the "submergence" of high-frequency spatial features. This visual degradation leads to vertical bounding box drift, which, as mathematically proven in Section 3.2, results in hyperbolic (inverse-square) distance estimation errors.

While LiDAR provides illumination-invariant 3D geometry, its high cost and hardware complexity restrict its use in mass-market edge devices, driving the need for purely monocular camera solutions [4].

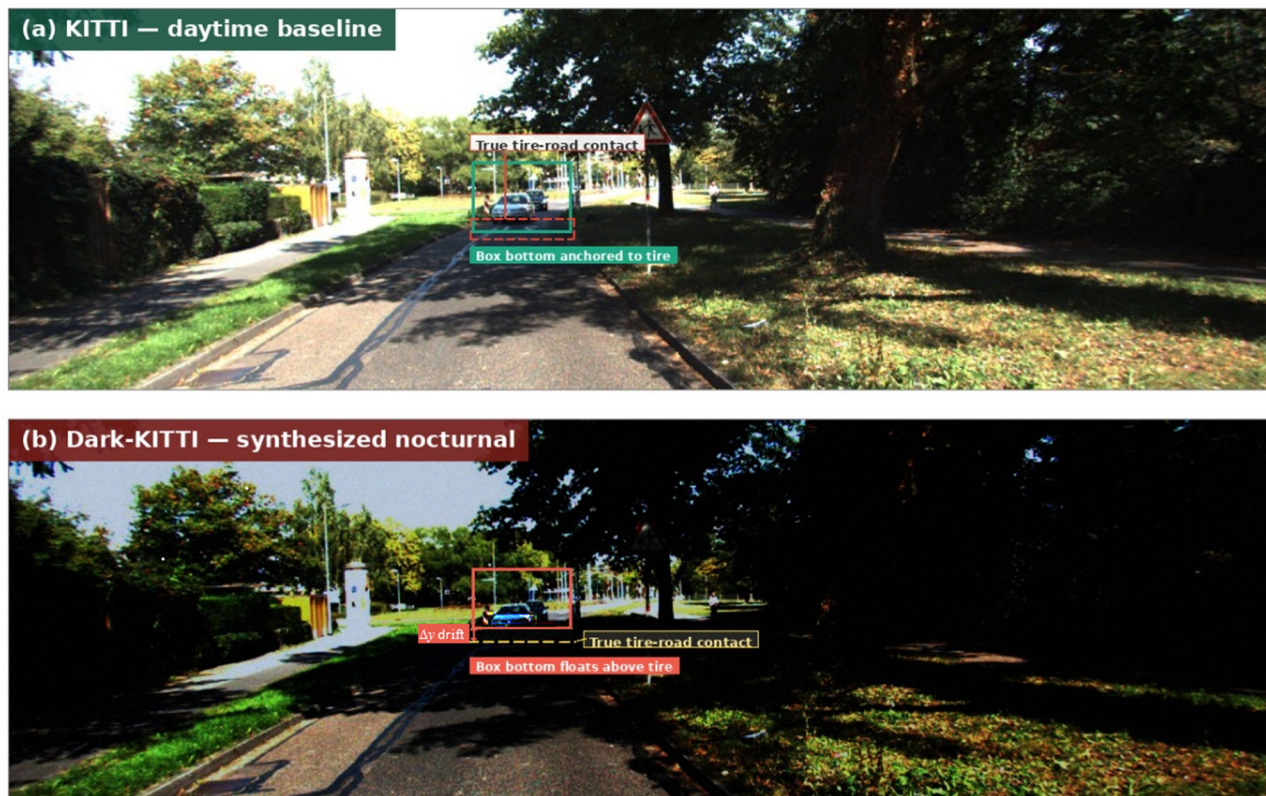


Figure 1. The Multi-Illumination Domain Gap. (a) KITTI daytime baseline: under high ambient illumination, the tire-road contact interface is sharply delineated (coral dashed region), and the predicted bounding box (teal) is correctly grounded to the true ground-contact pixel, enabling stable monocular pinhole projection. (b) Dark-KITTI synthesized nocturnal scene: the same tire-road interface is submerged into the elevated luminance noise floor; the predicted bounding box (coral, solid) “floats” above the true ground-truth contact (yellow dashed line) by Δy pixels, producing the multi-meter depth error mathematically derived in Section 3.2.

Recent SOTA monocular 3D detectors achieve high spatial accuracy in daylight conditions [5–8]. However, these models inherently rely on dense geometric feature extraction [9,10] and large model capacity, limiting their deployment on edge devices. Moreover, their performance degrades significantly in low-light scenarios due to the loss of high-frequency visual features [2]. In contrast, lightweight 2D detectors [11–13] provide efficient inference but lack geometric awareness. Their reliance on isotropic IoU-based loss functions [14] leads to vertical bounding-box drift, resulting in large depth estimation errors under the pinhole camera model [15].

To address these limitations, we propose the Anisotropic Geometry Loss (AGL) framework, a lightweight system for robust monocular distance estimation under cross-illumination conditions. The framework enforces geometric consistency through an anisotropic loss term—denoted $L_{y\text{tall}}$ —applied to the bounding-box bottom edge, and is further enhanced with a lightweight contrast normalization module for low-light environments. Throughout this paper, we use AGL to denote the trained detection framework (YOLOv10n + $L_{y\text{tall}}$ loss + decoupled head + safety clamp), $L_{y\text{tall}}$ to denote the loss term itself (defined in Section 3.3.1, combining Equations (5) and (6)), and AGL + CLAHE to denote the framework with the optional nocturnal preprocessor applied at inference time. The contributions of this work are summarized as follows:

1. We derive an inverse-square sensitivity analysis of the pinhole projection that mathematically motivates anisotropic bottom-edge weighting for monocular distance estimation, providing a principled alternative to isotropic IoU optimization.

2. We propose the $L_{y\text{tall}}$, a lightweight two-term penalty ($L_{\text{position}} + L_{\text{height}}$) that integrates into the YOLOv10n decoupled head without modifying the backbone, preserving real-time inference at >160 FPS on a 2.71 M-parameter footprint.
3. We introduce a luminance-channel CLAHE [16] preprocessor that recovers tire-road micro-contrast under nocturnal degradation [17], with asymmetric utility validated on synthetic Dark-KITTI and qualitatively corroborated on real ExDark imagery.
4. We further show that the two $L_{y\text{tall}}$ coefficients, λ_y and λ_h , which respectively govern vertical geometric alignment and object-scale consistency, exhibit strongly coupled optimization behavior rather than independently additive effects. This finding exposes a key limitation of conventional single-axis hyperparameter tuning. The apparent optimum of one coefficient shifts with the value of the other, potentially leading to misleading performance interpretations. To address this issue, we conduct a comprehensive joint sensitivity analysis to identify the true coupled optimum and establish a principled, reproducible guideline for coefficient selection in lightweight monocular 3D object detection.

Within these constraints, our experiments demonstrate that the AGL framework mitigates severe nocturnal domain shifts while maintaining inference speeds above 160 frames per second on highly constrained Nano architectures [18].

2. Related Work

2.1. Monocular 3D Object Detection

Monocular vision-based distance estimation has attracted significant attention due to its low hardware cost and ease of deployment. Recent advances in monocular depth estimation have focused on improving accuracy while maintaining computational efficiency for deployment in intelligent transportation systems. A recent survey by Huang et al. highlighted lightweight architectures, self-supervised learning, and robustness under challenging environmental conditions as major research directions in autonomous driving applications [19]. Modern monocular 3D detection models aim to recover lost depth information by imposing complex geometric constraints [20]. Architectures such as GUP-Net++ [5], MonoCon [8], and SMOKE [6] utilize sophisticated uncertainty prediction to achieve single-digit depth error on the standard KITTI dataset [21]. However, these models require upwards of 85 to 120 million parameters. This computational burden limits their deployment on standard AV edge hardware, and their reliance on high-fidelity visual textures makes them less robust when deployed across domains into low-light environments [22]. Therefore, we restrict the empirical comparison to lightweight 2D backbones operating at identical resolutions and with identical pinhole projections.

2.2. Advancements in 2D Architectures

Recent iterations of real-time object detectors, including YOLOv26 in an arXiv preprint [23], have introduced mechanisms such as Progressive Loss Balancing (ProgLoss) and Small-Target-Aware Label Assignment (STAL) to maximize 2D recall (mAP). While these innovations improve an architecture's ability to identify distant objects, they remain limited to 2D pixel space. Because they optimize for isotropic region overlap (IoU), they do not inherently address the monocular depth-projection failure caused by floating bounding boxes. Therefore, improving 2D IoU alone does not resolve 3D spatial errors [24]. Furthermore, recent architectures have heavily integrated attention mechanisms, ranging from Transformer-based architectures like DETR to attention-augmented detectors such as YOLOv12. Sui et al. developed a fusion-improved transformer framework for efficient monocular depth estimation [25], while Xia et al. demonstrated the effectiveness of pyramid transformers combined with multi-scale feature fusion for dense depth prediction [26].

While global self-attention improves occlusion handling by aggregating contextual image features, the localization heads of these architectures remain tethered to isotropic overlap metrics (e.g., GIoU or L1 loss). Consequently, attention-based models do not inherently encode 3D ground-plane awareness. They permit vertical bounding-box drift, which can degrade monocular pinhole distance projections.

2.3. Low-Light Perception

Low-light image enhancement remains an important research direction for improving perception reliability under adverse illumination conditions. Recent studies have utilized dynamic illuminance adjustment strategies to enhance image visibility and exposure consistency [27]. Efforts to improve nighttime perception typically fall into either sensor fusion or synthetic data generation [28]. Sensor fusion combines degraded camera feeds with LiDAR or Radar [29], violating the constraints of low-cost monocular systems [30]. Alternatively, Generative Adversarial Networks (GANs) [31] convert daytime images into synthetic nighttime data for training. While this augments datasets, it introduces generative hallucinations and does not mathematically fix the geometric failure of the bounding box itself [32]. Zhang and Lee introduced an exposure-aware diffusion framework to improve depth estimation robustness under adverse lighting conditions, demonstrating the continued importance of illumination-aware perception methods for autonomous driving systems [33].

Existing monocular 3D detection methods typically recover depth through dense feature extraction, keypoint estimation, uncertainty modeling, or attention-based contextual aggregation. While these approaches improve spatial reasoning under favorable illumination, they remain computationally intensive and continue to optimize localization primarily through isotropic overlap objectives. Consequently, they do not explicitly constrain the geometric stability of the tire–road contact boundary that governs monocular pinhole projection. In contrast, the proposed Anisotropic Geometry Loss (AGL) introduces an explicit bottom-edge-constrained anisotropic penalty derived from the inverse-square sensitivity of the pinhole camera model. Rather than increasing model capacity or attention complexity, AGL directly stabilizes the physically meaningful projection denominator responsible for depth estimation.

Geometry-aware learning has become an active research direction in monocular depth estimation. Xiong et al. incorporated geometric constraints into a self-supervised learning framework to improve depth consistency and estimation accuracy [34]. Such findings suggest that explicitly modeling geometric relationships can substantially improve depth prediction performance. Unlike these approaches, the proposed AGL focuses on directional distance estimation errors under low-light driving conditions.

Recent lightweight low-light image enhancement methods have demonstrated promising performance through channel-prior-guided enhancement and gamma correction mechanisms [35]. Such enhancement techniques can improve image visibility and feature representation, thereby benefiting downstream perception tasks such as object detection and distance estimation in low-light environments. Nevertheless, these approaches still require a dedicated learning-based inference stage. In contrast, CLAHE provides a simple and practical preprocessing solution without additional model parameters, making it suitable for resource-constrained deployment while improving local contrast prior to object detection and distance estimation. The framework is specifically designed for lightweight edge-scale detectors by demonstrating computational efficiency on a desktop GPU, enabling real-time cross-illumination distance estimation at >160 FPS while preserving a compact 2.71 M-parameter footprint.

3. Methodology

Figure 2 illustrates the proposed hierarchical system architecture of the AGL framework. The pipeline is partitioned into three hierarchical execution stages designed for computationally efficient deployment. Stage I utilizes CLAHE to restore micro-contrast at the object's base. Stage II employs a lightweight YOLOv10n backbone (2.3 M parameters) modified with an anisotropic y_{tall} regression penalty to enforce ground-plane stability. Stage III executes a pinhole camera projection and kinematic safety logic to classify objects within critical stopping distance zones (15 m/30 m) in accordance with Euro NCAP safety standards.

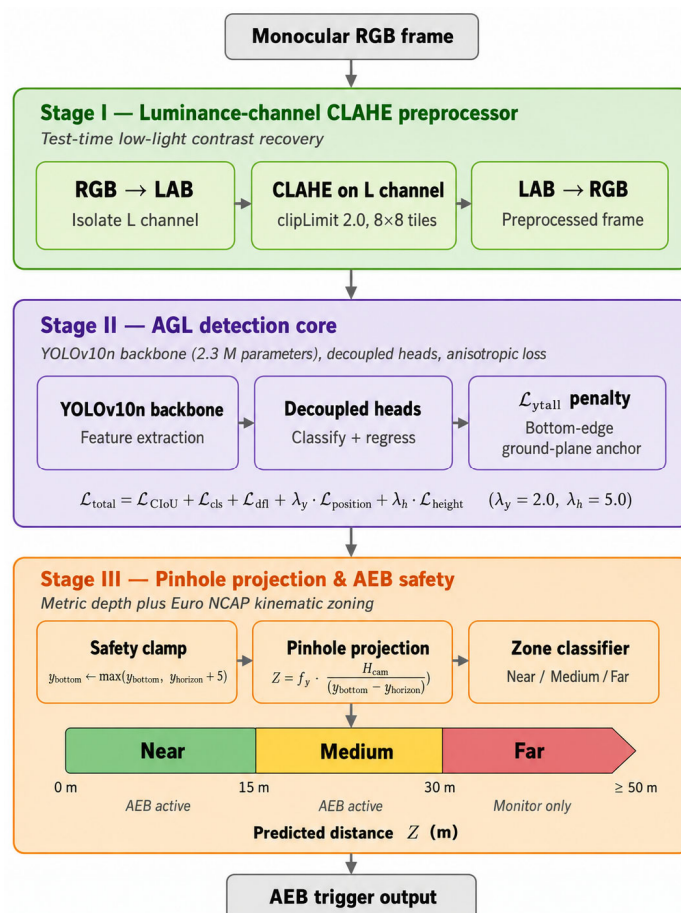


Figure 2. Overview of the proposed AGL system architecture, partitioned into three stages: CLAHE preprocessor (Stage I), AGL detection core (Stage II), and pinhole projection + AEB safety logic (Stage III).

3.1. Pinhole Geometry of Monocular Distance Recovery

Monocular distance estimation is fundamentally based on the pinhole camera model—a deterministic projective mapping from 3D world coordinates to 2D image pixels. Standard 2D object detectors operate exclusively in pixel space and discard the metric depth information implicit in this mapping. The image coordinate system follows the conventional computer vision convention, where the x-axis points rightward and the y-axis points downward. The central observation underlying our framework is that, under a small set of mild assumptions, depth can be recovered in closed form from a single geometrically stable pixel feature: the bottom edge of an object's bounding box. This section formalizes that recovery and identifies the sensitivity structure that motivates the loss design in Section 3.3.

Figure 3 illustrates the geometric setup. The camera is rigidly mounted at a fixed height H_{cam} above a planar ground surface, with its optical axis parallel to the ground

surface (zero pitch). The image row at which the ground horizon projects, y_{horizon} , divides the image into above-horizon and below-horizon regions. A target object whose tire-road contact line projects to image row y_{bottom} has its world depth Z fully determined by the camera intrinsics and the offset $(y_{\text{bottom}} - y_{\text{horizon}})$, as derived below.

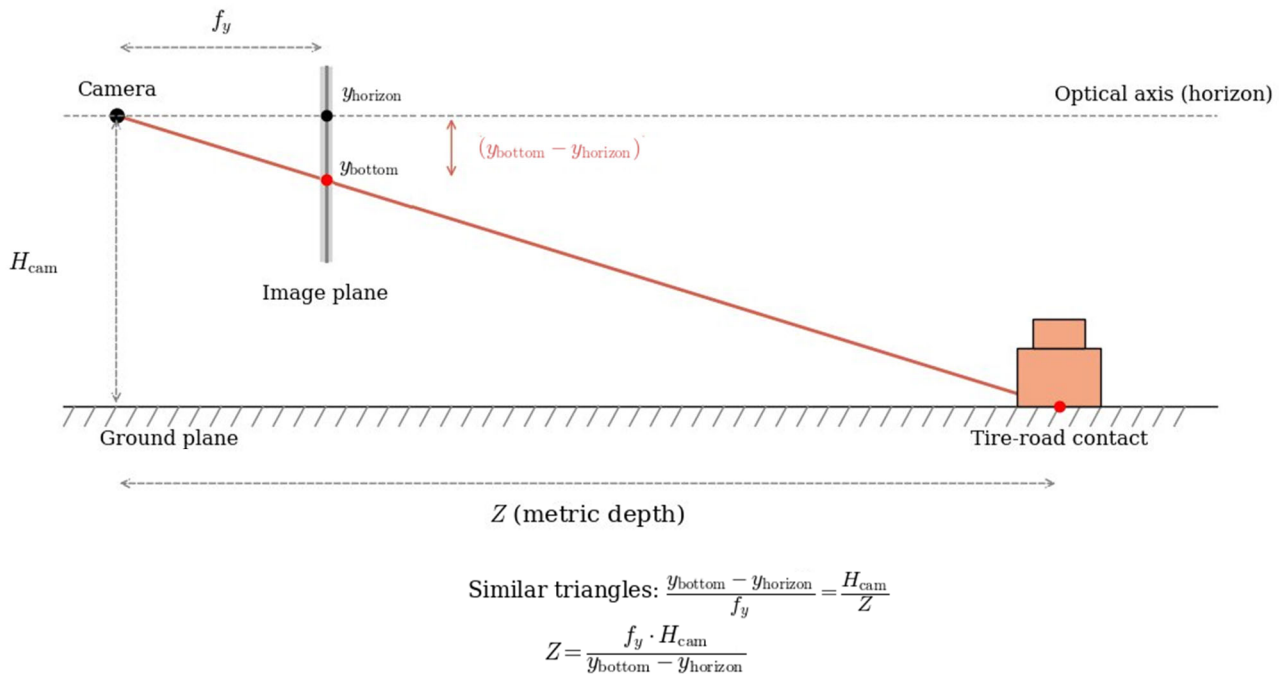


Figure 3. Pinhole geometry of monocular distance recovery under the flat-ground assumption. The camera is mounted at height H_{cam} above a planar ground surface with focal length f_y and principal-point row $c_y = y_{\text{horizon}}$. A target object's tire-road contact line projects to image row y_{bottom} , and its world depth Z is given by Equation (2) as a hyperbolic function of the pixel offset $(y_{\text{bottom}} - y_{\text{horizon}})$. The hyperbolic form implies that a unit-pixel error at the bottom edge translates into a metric depth error that scales as Z^2 , motivating the anisotropic bottom-edge penalty $L_{y_{\text{tall}}}$ introduced in Section 3.3.

Formally, the projection of a world point (X_w, Y_w, Z_w) to image pixel coordinates (x, y) is governed by the pinhole equation:

$$Z_c \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \end{bmatrix} \quad (1)$$

where f_x, f_y are the focal lengths in pixel dimensions, (c_x, c_y) is the principal point (the optical-center projection in image space; for KITTI, taken at the image center), and Z_c is the camera-frame depth—the z -component of the point's position in camera coordinates, which serves as the homogeneous-coordinate scale factor. Under the flat-ground assumption with the camera optical axis parallel to the ground plane, Z_c coincides with the world depth Z_w ; we therefore adopt the unsubscripted symbol Z throughout the remainder of the paper.

Placing the world origin at the camera's optical center, any point situated on the ground plane has a world Y -coordinate $Y_w = -H_{\text{cam}}$. Substituting this constraint into the v -row of Equation (1) and rearranging yields a closed-form expression for the depth of a ground-contact point in terms of a single observable, y_{bottom} :

$$Z = \frac{f_y \cdot H_{\text{cam}}}{y_{\text{bottom}} - y_{\text{horizon}}} \quad (2)$$

where y_{horizon} is the image row at which the optical vanishing point projects (equivalent to c_y under the zero-pitch assumption). Equation (2) is the geometric anchor of the entire framework: for any detected object whose bounding-box bottom edge corresponds to a point on the ground plane, the depth is determined by a single observable—the pixel offset ($y_{\text{bottom}} - y_{\text{horizon}}$)—scaled by the two calibration constants f_y and H_{cam} . Two consequences follow directly from this mapping and motivate the loss design in Section 3.3:

- **Inverse-square sensitivity.** Differentiating Equation (2) with respect to y_{bottom} shows that the depth error scales as $\Delta Z \propto Z^2 \cdot \Delta y / (f_y \cdot H_{\text{cam}})$; a unit pixel error at the bottom edge produces a metric depth error that grows quadratically with distance. This is the formal basis for the anisotropic bottom-edge weighting in the $L_{y\text{tall}}$ loss.
- **Bottom-edge specialization.** The recovery depends exclusively on the bounding box's bottom edge, not its center, width, or top edge. Standard isotropic IoU-based regression treats all four box edges as equivalent, dispersing supervision across geometrically uninformative coordinates. $L_{y\text{tall}}$ concentrates supervision on the single edge that carries the metric depth signal.

The flat-ground and zero-pitch assumptions underlying Equation (2) merit explicit treatment. On graded terrain or during dynamic vehicle motion (braking, acceleration, suspension travel), a non-zero pitch angle ψ introduces an additive bias to y_{horizon} of approximately $f_y \cdot \tan(\psi)$, which propagates into the depth estimate via the inverse-square sensitivity noted above. At the KITTI calibration values ($f_y = 721.53$ px, $H_{\text{cam}} = 1.65$ m), a pitch deviation of 1° produces a horizon shift of approximately 12.6 pixels and a relative depth error of approximately 31% at $Z = 30$ m. The KITTI recording vehicle operates predominantly on level urban roads, validating the static-horizon assumption for the numerical results reported throughout this paper. Extension of the framework to graded terrain via lightweight dynamic pose estimation is reserved for future work and discussed in Section 7.

3.2. The Non-Linear Depth Decay and the Necessity of $L_{y\text{tall}}$

Standard object detectors use IoU-based loss functions [14], which treat bounding-box errors isotropically. Unlike standard IoU-based loss functions that assume a linear relationship between bounding box pixel error and spatial error, the proposed $L_{y\text{tall}}$ anisotropic penalty explicitly maps the non-linear monocular projection geometry. Taking the first derivative of the depth equation with respect to y_{bottom} reveals the true, non-linear hyperbola of monocular depth:

$$\frac{dZ}{dy_{\text{bottom}}} = -\frac{f_y \cdot H_{\text{cam}}}{(y_{\text{bottom}} - y_{\text{horizon}})^2} \quad (3)$$

The derivative shows that depth sensitivity increases hyperbolically (as an inverse-square function) as the bounding box prediction (y_{bottom}) approaches the optical horizon (y_{horizon}). A 5-pixel error at the bottom of the image (near range) results in a centimeter-scale depth error. Conversely, the same 5-pixel error near the horizon (far range) results in a multi-meter depth error of dozens of meters.

Standard 2D detectors like YOLOv10n optimize for IoU, allowing the bottom edge of the box to “float” a few pixels above the tires without penalty. While acceptable for 2D classification, the mathematical proof above demonstrates why this floating behavior causes large multi-meter distance errors. To counteract this, AGL introduces the y_{tall} Anisotropic Penalty, forcing the network to explicitly minimize the squared bottom-edge error $(y_{\text{bottom}} - y_{\text{horizon}})^2$, ensuring the structural integrity of the denominator in our depth projection.

To prevent numerical instability (division by zero) when visual noise causes a bounding box to cross the optical horizon, we dynamically enforce a spatial safety clamp:

$$y_{\text{bottom}} = \max(y_{\text{bottom}}, y_{\text{horizon}} + 5.0) \quad (4)$$

At the KITTI calibration values ($f_y = 721.53$, $H_{\text{cam}} = 1.65$ m), this +5.0 pixel buffer corresponds to a maximum detectable range of $Z = 721.53 \times 1.65/5 \approx 238$ m—well beyond the Euro NCAP’s 30 m far boundary—ensuring numerical stability without truncating any safety-relevant detection. This +5.0 pixel buffer represents the practical limit of the sensor’s far-range visibility, ensuring safe and stable computations within the Autonomous Emergency Braking (AEB) kinematic logic.

3.3. Hierarchical Loss Function Design

The key contribution of AGL lies in modifying the objective function used during backpropagation. Standard 2D detectors use isotropic loss functions that treat vertical and horizontal bounding-box errors equally. However, as proven in Section 3.2, monocular depth estimation is uniquely sensitive to vertical drift. To resolve this, we propose a Hierarchical Loss Architecture that integrates our custom geometric penalty into the modern decoupled head of the YOLOv10n backbone [13]. YOLOv10n was selected precisely because it eliminates Non-Maximum Suppression (NMS), guaranteeing consistent inference speeds on edge hardware without post-processing latency bottlenecks.

Furthermore, standard object detectors historically relied on shared “sibling heads” for both classification and localization. However, these objectives are inherently misaligned; classification requires translation invariance to recognize objects, while localization requires translation variance to regress the bounding box coordinates precisely [36]. Modern architectures address this via decoupled heads [37,38]. This architectural decoupling is critical to AGL, as it allows the anisotropic penalty to be applied exclusively to the bounding-box regression head. By strictly optimizing the localization head for 3D ground-plane adherence, we force the network to learn geometric constraints without compromising the classification head’s 2D recall.

3.3.1. The y_{tall} Anisotropic Ground-Plane Penalty

We define the positional loss L_{position} as an anisotropic penalty that specifically targets the lower boundary of the predicted bounding box. Unlike the standard IoU loss, which optimizes the entire area of the box, L_{position} acts as a spatial anchor. It ensures that the predicted ground-contact point $y_{\text{pred_bottom}}$ converges toward the true tire-asphalt interface $y_{\text{target_bottom}}$.

We utilize the L_2 -norm (Squared Error) for this penalty to generate an increasingly aggressive gradient as the box “floats” away from the road surface:

$$L_{\text{position}} = (y_{\text{pred_bottom}} - y_{\text{target_bottom}})^2 \quad (5)$$

However, penalizing only the bottom-edge position is insufficient in practice, as the bottom edge is co-determined by the bounding box vertical center and its height: $y_{\text{bottom}} = y_{\text{center}} + h/2$. A network optimizing solely on Equation (5) may satisfy the constraint by inflating the predicted bounding box height while leaving the true ground-contact pixel uncovered, thereby undermining the geometric anchoring objective. To prevent this compensatory behavior, we introduce a complementary height regularization term that penalizes deviation between predicted and target bounding box heights:

$$L_{\text{height}} = (h_{\text{pred}} - h_{\text{target}})^2 \quad (6)$$

The positional term penalizes the vertical displacement of the predicted bottom edge: $L_{\text{position}} = (y_{\text{pred_bottom}} - y_{\text{target_bottom}})^2$. The full tall penalty combines both terms: $L_{y\text{tall}} = \lambda_y \cdot L_{\text{position}} + \lambda_h \cdot L_{\text{height}}$, where λ_y and λ_h are the weighting coefficients introduced in Section 3.3.2. The decomposition explicitly separates positional and scale-related constraints: λ_y controls how strongly the network is pulled toward the correct bottom-edge position. At the same time, λ_h prevents the network from satisfying the position constraint via height inflation. The empirical setting of these coefficients ($\lambda_y = 2.0$, $\lambda_h = 5.0$) is justified in Section 4.5 below.

3.3.2. Total Multi-Task Objective Function

To ensure that the network maintains high 2D detection performance while gaining 3D awareness, the $L_{y\text{tall}}$ penalty is integrated into a multi-task loss formulation. The total objective function L_{total} is defined as follows:

$$L_{\text{total}} = \lambda_{\text{box}} L_{\text{CIoU}} + \lambda_{\text{cls}} L_{\text{cls}} + \lambda_{\text{dfl}} L_{\text{dfl}} + \lambda_y L_{y\text{tall}} \quad (7)$$

where each λ is a scalar weighting hyperparameter that balances the constituent loss terms. L_{CIoU} (Complete-IoU Loss) optimizes the overlap, aspect ratio, and center-distance of the bounding box. L_{cls} (Classification Loss) uses Binary Cross-Entropy to ensure correct category identification (e.g., Car vs. Pedestrian). L_{dfl} (Distribution Focal Loss) optimizes the probability distribution of the box boundaries for sub-pixel accuracy. $L_{y\text{tall}}$ (AGL Innovation) enforces the geometric ground-plane constraint, implemented as the two-term penalty $L_{y\text{tall}} = \lambda_y \cdot L_{\text{position}} + \lambda_h \cdot L_{\text{height}}$ defined in Section 3.3.1, with $\lambda_y = 2.0$ and $\lambda_h = 5.0$.

The coefficient λ_y is a critical hyperparameter that balances the geometric constraint against traditional 2D detection goals. If λ_y is too low, the model defaults to standard 2D behavior with high depth error; if too high, the model may experience bounding box collapse, where it prioritizes the bottom edge so much that it loses track of the object's height and width. As empirically demonstrated in Section 5.4, increasing λ_y beyond the optimal regime results in measurable degradation in RMSE, providing quantitative evidence of the trade-off between geometric constraint and bounding box stability. The companion coefficient λ_h weights the height-regularization term L_{height} (Equation (6)) and counteracts the network's tendency to satisfy the position constraint via height inflation; in our experiments, λ_h is held at 5.0 throughout.

3.4. CLAHE Luminance-Channel Preprocessor

By converting the RGB input to the LAB color space, we isolate the Luminance (L) channel and apply CLAHE. This amplifies the micro-contrast between the dark vehicle tire and the dark asphalt without globally overexposing the image. We acknowledge a physical limitation in environments with specular surfaces, such as rain-slicked nocturnal asphalt or wet road markings, where headlight reflections produce localized regions of saturated luminance. In such cases, the CLAHE histogram redistribution may amplify the specular highlight rather than the useful tire-road micro-contrast, potentially injecting noise into the y_{tall} gradient. Quantitative evaluation under these conditions requires a dedicated wet-surface nocturnal benchmark and is reserved for future investigation alongside the pose-compensation extension of Section 7.

4. Experimental Setup

To ensure rigorous reproducibility of the AGL framework and its benchmarks, the following section details the dataset synthesis, hardware environment, and algorithmic configurations used in our experiments.

4.1. Dataset Composition: KITTI and Dark-KITTI Synthesis

The framework was calibrated and trained from scratch using the widely used KITTI benchmark. The standard KITTI dataset contains 7481 training images and 7518 test images captured from a moving vehicle in a mid-size city, providing highly accurate LiDAR-derived 3D ground truth labels for five target categories: Car, Van, Truck, Pedestrian, and Cyclist.

However, because KITTI is composed exclusively of daytime, high-visibility scenarios, testing the nocturnal domain shift required the algorithmic generation of the Dark-KITTI validation set used in this work. Dark-KITTI was generated by applying a low-light degradation pipeline to the standard KITTI validation split. This pipeline included non-linear gamma correction ($\gamma = 2.5$) to simulate low ambient lighting, the injection of Poisson-Gaussian noise to emulate the low signal-to-noise ratio of night-vision sensors, and localized overexposure masking to simulate headlight glare. This resulted in a physically consistent multi-illumination evaluation dataset preserving the exact 3D LiDAR ground truths of the original dataset. For zero-shot domain generalization tests, we utilized the ExDark (Exclusively Dark) dataset [39], which contains real-world low-light imagery captured from diverse, uncalibrated camera perspectives.

4.2. Hardware and Software Environment

All training, ablation studies, and live-inference evaluations were conducted on a local simulation workstation equipped with a single NVIDIA RTX 4070 GPU (12 GB VRAM), an Intel Core i7 processor, and 32 GB of RAM. The software stack was built on Python 3.10 and PyTorch 2.0, using a heavily customized fork of the Ultralytics framework to accommodate the decoupled $L_{y_{\text{tall}}}$ regression head and the real-time AEB logic triggers. Parameter counts are reported under two conventions: the YOLOv10n backbone alone is 2.3 M parameters (per the original paper [13], measuring the fused inference graph of the feature-extraction backbone), while the full trainable model, including the neck and decoupled head used in this work, is 2.71 M parameters (Table 1). Figure 2 references the backbone count to describe the architectural starting point; Table 1 reports the full count for cross-model comparison consistency.

Table 1. Cross-domain localization performance on KITTI-Bright (day) and Dark-KITTI (night). **Bold** = best; *italics* = second-best; underline = third-best. Lower RMSE is better. All AGL configurations use ($\lambda_y = 2.0$, $\lambda_h = 5.0$); FPS measured on NVIDIA RTX 4070 at 640×480 input.

Method	Params (M)	Inference (FPS)	KITTI-Bright (Day) RMSE (m)	Dark-KITTI (Night) RMSE (m)
Vanilla YOLOv10n [13]	<u>2.71</u>	194	11.12 ± 0.53	11.53 ± 0.56
Vanilla YOLOv26n [23]	2.51	147	11.08 ± 0.65	11.99 ± 0.58
AGL (Ours)	<u>2.71</u>	<u>177</u>	10.31 ± 0.58	<u>10.91 ± 0.68</u>
AGL + CLAHE	<u>2.71</u>	176	<u>10.52 ± 0.56</u>	10.55 ± 0.72

Note: A paired Wilcoxon signed-rank testing on matched detections confirmed that the Dark-KITTI improvements are statistically significant relative to the vanilla YOLOv10n baseline for both AGL ($p = 2.55 \times 10^{-11}$) and AGL + CLAHE ($p = 7.51 \times 10^{-22}$) at the significance level $\alpha = 0.05$.

All FPS values reported throughout this paper (Tables 1–3) are obtained from a single unified measurement session conducted in parallel with the generation of the per-detection (Z_{pred} , Z_{gt}) pairs that feed the bootstrap confidence-interval analysis described in Section 4.6. This unified protocol ensures methodological consistency across tables—a given configuration produces the same reported FPS whether it appears in Tables 1 and 3, or elsewhere. Inter-session benchmark variance on the same RTX 4070 hardware and model can reach approximately 5–15% depending on GPU thermal and cache state at measurement time.

Table 2. Spatial distance decay on Dark-KITTI (intersection set, IoU ≥ 0.5). The Near zone (0–15 m), Medium zone (15–30 m), and Far zone (>30 m) correspond to Euro NCAP AEB kinematic stopping requirements for urban (30 km/h) and suburban (50 km/h) operation. **Bold** = best in row group. Lower is better for RMSE/MAE/Abs Rel; higher is better for δ_1 .

Distance Range (m)	Model	RMSE (m)	MAE (m)	Abs Rel	δ_1	N
Near (0–15)	Vanilla YOLOv26n	2.42 $\pm$ 0.03	2.16 $\pm$ 0.03	0.264 $\pm$ 0.008	43.8 $\pm$ 1.4%	1250
	Vanilla YOLOv10n	2.43 $\pm$ 0.03	2.16 $\pm$ 0.03	0.264 $\pm$ 0.008	44.2 $\pm$ 1.4%	1250
	AGL (Ours)	2.42 $\pm$ 0.03	2.16 $\pm$ 0.03	0.264 $\pm$ 0.008	44.1 $\pm$ 1.3%	1250
Medium (15–30)	Vanilla YOLOv26n	5.96 $\pm$ 0.18	5.23 $\pm$ 0.07	0.232 $\pm$ 0.003	34.8 $\pm$ 1.2%	1590
	Vanilla YOLOv10n	6.12 $\pm$ 0.25	5.24 $\pm$ 0.08	0.233 $\pm$ 0.003	34.1 $\pm$ 1.2%	1590
	AGL (Ours)	5.94 $\pm$ 0.19	5.15 $\pm$ 0.07	0.229 $\pm$ 0.003	35.7 $\pm$ 1.2%	1590
Far (>30)	Vanilla YOLOv26n	17.59 $\pm$ 1.25	13.59 $\pm$ 0.29	0.314 $\pm$ 0.006	16.8 $\pm$ 0.9%	1438
	Vanilla YOLOv10n	17.41 $\pm$ 1.23	13.52 $\pm$ 0.29	0.312 $\pm$ 0.006	16.8 $\pm$ 1.0%	1438
	AGL (Ours)	17.41 $\pm$ 1.25	13.45 $\pm$ 0.29	0.311 $\pm$ 0.006	17.3 $\pm$ 1.0%	1438

Note: Tabulated values are point estimate $\pm$ bootstrap standard deviation (1000 iterations). All three models evaluated on an identical ground-truth detection set to eliminate differential recall as a confound.

Table 3. Resolution ablation on Dark-KITTI. Five configurations spanning two input resolutions (1280 px, 640 px) and three architectural variants. Lower is better for RMSE/MAE/Abs Rel/Memory; higher is better for δ_1 and FPS. **Bold** = best in column; *italics* = second-best. GFLOPs (billions of floating-point operations) measure the computational workload required for one inference. Higher GFLOPs indicate greater computational complexity and typically correspond to lower inference speed (FPS), reflecting the trade-off between accuracy and computational efficiency.

Model	RMSE (m)	MAE (m)	Abs Rel	δ_1	GFLOPs	FPS	Memory (MB)	N
Vanilla YOLOv10n @1280	12.14 $\pm$ 0.56	8.10 $\pm$ 0.13	0.279 $\pm$ 0.003	31.5 $\pm$ 0.6%	32.8	130	118	5292
AGL @1280	10.92 $\pm$ 0.52	7.59 $\pm$ 0.12	0.264 $\pm$ 0.003	31.5 $\pm$ 0.7%	32.8	110	131	4753
AGL + CLAHE @1280	10.34 $\pm$ 0.42	7.39 $\pm$ 0.11	0.258 $\pm$ 0.002	31.9 $\pm$ 0.7%	32.8	117	139	4458
AGL @640	10.91 $\pm$ 0.68	7.15 $\pm$ 0.13	0.268 $\pm$ 0.003	32.1 $\pm$ 0.7%	8.2	182	87	4461
AGL + CLAHE @640	10.55 $\pm$ 0.72	6.97 $\pm$ 0.12	0.260 $\pm$ 0.003	33.3 $\pm$ 0.7%	8.2	164	95	4311

4.3. Training Configurations and Hyperparameters

Both the baseline YOLOv10n and the AGL architectures were trained for 100 epochs. Preliminary experiments indicated that the validation mAP@0.5 converged by approximately epoch 80, with negligible improvement beyond epoch 100. The cosine annealing schedule reaches its terminal learning rate by epoch 100, ensuring stable final-state weights across all sweep configurations within a tractable computational budget. We utilized the AdamW optimizer, with an initial learning rate of 0.001, momentum of 0.937, and a cosine annealing learning rate scheduler.

To demonstrate that algorithmic efficiency supersedes brute-force hardware scaling, the standard input resolution for AGL was set to 640×480 pixels (maintaining the aspect ratio via proportional letterboxing). This constraint was critical for maintaining the >160 FPS inference speed required for real-time vehicular actuation.

4.4. Lightweight Enhancement Filter Parameters

The test-time lightweight enhancement filter was implemented using OpenCV's CLAHE algorithms. The raw RGB frames were converted to the LAB color space to isolate structural illumination. The CLAHE algorithm was applied strictly to the Luminance (L) channel using a clipLimit of 2.0 and a tileGridSize of (8×8) . These specific parameters were empirically chosen to prevent global noise amplification while maximizing the high-frequency micro-contrast at the critical tire-road boundary, providing the exact pixel gradients required by the y_{tall} penalty in low-light environments.

4.5. Distance Estimation and Inference

To ensure metric accuracy, AGL utilizes the official intrinsic and extrinsic hardware calibrations of the KITTI research vehicle [21]. The vertical focal length of the left color camera f_y is set to 721.53 pixels, and the optical horizon center y_{horizon} is located at pixel row 165.0. The physical height of the camera lens above the asphalt H_{cam} is fixed at 1.65 m. Because Z is inversely proportional to y_{bottom} , a 2D bounding box that “floats” even a few pixels above the true 0.0 m tire-road contact point will result in hyperbolic (inverse-square) distance calculation errors.

During inference, the spatial safety constraint $y_{\text{bottom}} = \max(y_{\text{bottom}}, y_{\text{horizon}} + 5.0)$ was applied. This buffer prevents division-by-zero errors at the vanishing point. Without this clamp, nocturnal noise that causes boxes to overlap the horizon line would result in infinite-distance predictions, destabilizing the system’s safety-triggering mechanisms.

For the AGL framework, λ_y is selected as a stability-oriented configuration set to 2.0, paired with a height-regularization coefficient $\lambda_h = 5.0$ (introduced in Section 3.3.1). These coefficients reflect the gradient-mechanics analysis in Section 3.3.2 and are selected within an empirically validated bounded regime. Within this regime, AGL behavior is stable: the framework retains positive RMSE improvement over the Vanilla YOLOv10n baseline across all five tested λ_y values (0.5, 1.0, 2.0, 3.0, 5.0) and all three tested λ_h values (2.0, 5.0, 10.0), supporting practical hyperparameter selection in deployment without requiring exhaustive per-application tuning. This weighting prioritizes geometric ground-plane consistency during training. As training progresses, the gradients from $L_{y\text{tall}}$ dominate y -axis refinement, anchoring the model to the road surface and suppressing the rapid error amplification previously discussed.

4.6. Evaluation Metrics

To quantify the spatial localization accuracy of AGL under both diurnal and nocturnal operating conditions, we adopt four complementary metrics drawn from the standard monocular depth estimation literature. Each metric captures a distinct failure mode, and their combined interpretation provides a comprehensive view of both the average and worst-case behavior that no single metric can provide in isolation. Let N denote the number of validated detections, $Z_{\text{pred},i}$ the predicted metric depth of the i -th sample, and $Z_{\text{gt},i}$ the corresponding LiDAR-derived ground-truth distance taken from the KITTI 3D object labels.

Following standard KITTI practice, the distance-estimation metrics (RMSE, MAE) are reported on the Car class, which provides the most numerous and best-calibrated 3D ground-truth instances for statistically robust evaluation. The AGL framework is trained on the full 5-class KITTI label set (Car, Van, Truck, Pedestrian, Cyclist) for representational diversity and detects all classes during inference. Qualitative cross-class generalization on Pedestrian and Cyclist categories under real-world low-light conditions is demonstrated on the ExDark dataset (Section 5.3.2). Predictions are matched to ground-truth instances via $\text{IoU} \geq 0.5$ prior to metric computation.

Root Mean Square Error (RMSE). RMSE is the primary reporting metric throughout this work and is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (Z_{\text{pred},i} - Z_{\text{gt},i})^2} \quad (8)$$

Because RMSE squares each residual before averaging, it disproportionately penalizes large deviations. This property is essential in autonomous driving: a single severe outlier—such as the floating bounding box failure mode derived in Section 3.2—will dominate the metric and expose safety-critical failures that linear measures would mask. RMSE is reported in meters throughout Tables 1–3, with lower values indicating better performance.

Mean Absolute Error (MAE). MAE complements RMSE by providing a linear, scale-faithful measure of the typical deviation:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |Z_{\text{pred},i} - Z_{\text{gt},i}| \quad (9)$$

Because MAE weights all residuals equally, it is more representative of the depth error perceived on an “average” frame. It is therefore the metric most directly interpretable against the Euro NCAP kinematic stopping boundaries [40]. Joint reporting of RMSE and MAE allows the reader to distinguish between models that fail occasionally but significantly (where RMSE substantially exceeds MAE) and those that fail consistently but mildly (where the two metrics converge). MAE is reported in meters.

Absolute Relative Error (Abs Rel). Because monocular depth decays hyperbolically with y_{bottom} per Equation (3), a 2 m error observed at 10 m range carries a fundamentally different physical meaning than a 2 m error at 50 m range. To normalize for this distance-dependent sensitivity, we report the Absolute Relative Error:

$$\text{Abs Rel} = \frac{1}{N} \sum_{i=1}^N \frac{|Z_{\text{pred},i} - Z_{\text{gt},i}|}{Z_{\text{gt},i}} \quad (10)$$

This dimensionless metric rescales each residual by its ground-truth distance, isolating the geometric quality of the prediction from the absolute depth magnitude. Abs Rel is particularly diagnostic in the Far range (>30 m), where both RMSE and MAE inflate simply as a consequence of the inverse-square projection geometry rather than any deterioration of the model’s underlying ground-plane adherence.

Threshold Accuracy (δ_1). Finally, we report threshold accuracy at the canonical $\delta_1 = 1.25$ level established by Eigen et al. [40]:

$$\delta_1 = \frac{1}{N} \sum_{i=1}^N \mathbb{1} \left(\max \left(\frac{Z_{\text{pred},i}}{Z_{\text{gt},i}}, \frac{Z_{\text{gt},i}}{Z_{\text{pred},i}} \right) < 1.25 \right) \quad (11)$$

δ_1 quantifies the fraction of predictions whose multiplicative ratio to the ground truth lies within a $\pm 25\%$ window, and it has been adopted near-universally in the monocular depth literature since its introduction. Unlike the three error metrics above, δ_1 is an accuracy metric, with higher values indicating better performance. Reporting δ_1 alongside RMSE, MAE, and Abs Rel allows us to verify that reductions in error are not achieved at the cost of a collapse in the proportion of correctly localized samples—a failure mode that the geometric weighting coefficient λ_y is particularly prone to triggering when set too aggressively, as discussed in Section 3.3.2.

Following the stratified evaluation protocol of [41], all four metrics are reported both as aggregates over the full detection set (Tables 1 and 3) and as stratifications by kinematic safety zone—Near (0–15 m), Medium (15–30 m), and Far (>30 m)—to enable the zone-wise decay analysis reported in Section 5.2.

To quantify statistical uncertainty in tabulated RMSE, MAE, Abs Rel, and δ_1 values arising from finite validation sample size, we report 95% bootstrap confidence intervals computed via 1000-iteration resampling with replacement of the per-detection $(Z_{\text{pred}}, Z_{\text{gt}})$ pairs [42]. Tabulated values throughout the paper are reported as point estimate $\pm$ bootstrap standard deviation. This convention captures finite-sample statistical uncertainty separately from training stochasticity, the latter addressed via the seed-sensitivity analysis in Section 5.4 within the scope constraints described in Section 6.6.

To enhance methodological reproducibility, all implementation details, coefficient configurations, preprocessing procedures, and evaluation protocols are explicitly docu-

mented. The Dark-KITTI synthesis pipeline parameters ($\gamma = 2.5$, Poisson-Gaussian noise specifications, overexposure mask configuration), CLAHE parameters (clipLimit = 2.0, tileGridSize 8×8 , LAB color space), KITTI calibration values ($f_y = 721.53$ px, $H_{\text{cam}} = 1.65$ m, $y_{\text{horizon}} = 165$ px), and AGL training hyperparameters (AdamW optimizer, lr = 0.001, momentum = 0.937, cosine annealing, 100 epochs, $\lambda_y = 2.0$, $\lambda_h = 5.0$) are fully specified in Sections 4.1–4.5. The source code and pretrained models supporting the findings of this study are available from the corresponding author upon reasonable request.

5. Experimental Results and Discussion

5.1. Cross-Domain Robustness and Model Scale

In the baseline model, the absence of geometric constraints allows the bounding box to “float” above the true road surface, resulting in a distance error that violates kinematic safety margins. In contrast, the proposed AGL framework uses the $L_{y_{\text{tall}}}$ penalty to constrain the prediction to the tire-asphalt interface, delivering precise metric localization even in severe low-light conditions, as shown in Figure 4.

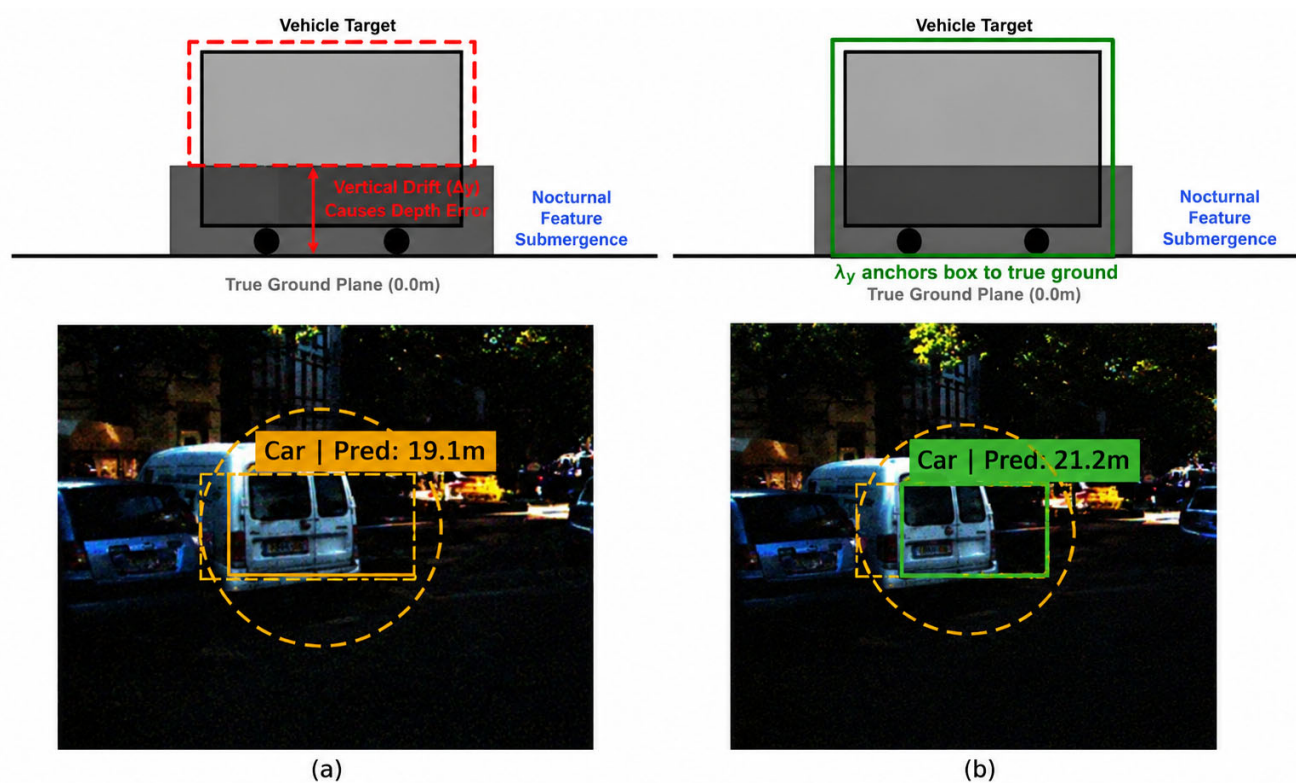


Figure 4. Qualitative comparison of spatial anchoring conceptual sketch (top) and the real inference results (bottom) on a Dark-KITTI frame (Medium zone, ground-truth Car at $Z = 22.49$ m). The yellow dashed rectangle marks the KITTI LiDAR-derived 3D label, and the coral dashed circle highlights the target of interest. (a) Vanilla YOLOv10n predicts $Z = 19.10$ m, underestimating depth by 3.38 m (15.0% relative error). In the absence of a ground-plane constraint, the bottom edge of the predicted bounding box drifts from the true tire-asphalt interface, and the inverse-square pinhole sensitivity derived in Equation (3) amplifies the small pixel drift into a multi-meter metric error. (b) AGL with the CLAHE filter predicts $Z = 21.20$ m on the identical target ($|\text{err}| = 1.29$ m, 5.7% relative error), reducing the absolute depth error by 62% through y_{tall} 's anisotropic penalty on the bounding-box bottom edge. Each panel's upper-right inset displays the model's predicted metric depth (m) for the highlighted target object.

Table 1 quantifies cross-domain localization across the lightweight 2D detector landscape. The Vanilla YOLOv26n baseline (2.51 M parameters), despite its newer architectural

design, including attention mechanisms and Progressive Loss Balancing, achieves the worst RMSE across both domains (11.08 ± 0.65 m in daylight, 11.99 ± 0.58 m in nocturnal). This is consistent with our hypothesis from Section 2.2 that 2D architectural improvements alone do not address the geometric failure of monocular depth projection. Vanilla YOLOv10n [13] performs marginally better (11.12 ± 0.53 m daylight, 11.53 ± 0.56 m nocturnal) due to its NMS-free decoupled head, but still suffers from the underlying isotropic-loss limitation.

By introducing the y_{tail} anisotropic penalty, AGL reduces daylight RMSE to 10.31 ± 0.58 m (-0.81 m vs. YOLOv10n; -0.77 m vs. YOLOv26n) while preserving real-time inference at 177 FPS. AGL alone improves nocturnal RMSE to 10.91 ± 0.68 m. The framework's specialized cross-illumination configuration (AGL + CLAHE) achieves the best nocturnal performance at 10.55 ± 0.72 m, a 0.98 m improvement over Vanilla YOLOv10n and a 1.44 m improvement over Vanilla YOLOv26n.

CLAHE exhibits asymmetric behavior consistent with its design rationale: evaluation of AGL + CLAHE on KITTI-Bright (10.52 ± 0.56 m) showed marginal RMSE degradation relative to AGL alone (10.31 ± 0.58 m), reflecting the loss of acuity from histogram equalization in already-saturated illumination. During nocturnal evaluation, RMSE improves ($10.91 \pm 0.68 \rightarrow 10.55 \pm 0.72$ m) because the same operator amplifies the tire-road micro-contrast suppressed by darkness [32]. This validates CLAHE's specialized utility as a zero-shot nocturnal safety trigger rather than a universal preprocessor. Notably, AGL + CLAHE achieves this nocturnal RMSE improvement at essentially the same inference speed as AGL alone (176 vs. 177 FPS), confirming that CLAHE adds negligible end-to-end latency.

To visualize the cost–capability trade-off underlying Table 1, Figure 5 plots the Dark-KITTI RMSE of every evaluated configuration against its inference speed. The visualization makes the Pareto structure of the comparison explicit: both AGL configurations achieve lower RMSE than either Vanilla baseline while operating well within the >140 FPS AEB real-time zone, suggesting that the AGL configurations provide a more favorable accuracy–speed trade-off than the Vanilla baselines.

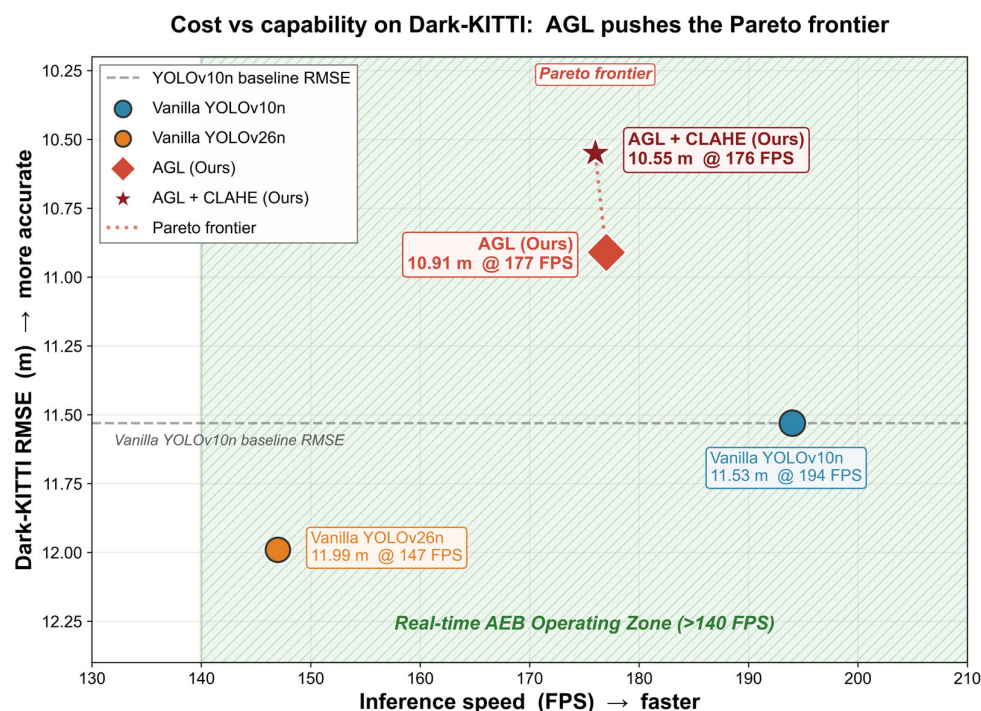


Figure 5. Cost vs. capability on Dark-KITTI. Each method is plotted by its inference speed (x -axis, higher = faster) against its depth-estimation RMSE (y -axis, lower = more accurate; axis inverted so

“better” is up). Marker shape distinguishes baseline methods (○) from our methods (◇ AGL, ★ AGL + CLAHE). The dashed gray line marks the Vanilla YOLOv10n baseline RMSE for reference; the red dotted line connects the two Pareto-optimal configurations in our search space. Both AGL configurations Pareto-dominate the Vanilla baselines, and all four methods operate within the >140 FPS real-time AEB zone (green shaded).

5.2. Kinematic Distance Decay Analysis

To strictly evaluate the monocular distance estimation within the context of vehicle safety, the spatial evaluation was partitioned into three distinct operational zones: Near (0–15 m), Medium (15–30 m), and Far (>30 m). These boundaries were selected to reflect the kinematic stopping distance requirements defined by the European New Car Assessment Programme (Euro NCAP) for Autonomous Emergency Braking (AEB) systems [41]. Specifically, the 15 m boundary represents the critical stopping distance required for an urban vehicle traveling at 30 km/h, while the 30 m boundary reflects the stopping constraints at 50 km/h [43].

Figure 6 shows the 15-m (amber) and 30-m (coral) thresholds, representing the approximate stopping-distance thresholds required for urban vehicles traveling at 30 km/h and 50 km/h, respectively. This visualization demonstrates how AGL prioritizes precision in the “Near Zone” (0–15 m), where depth estimation errors are most likely to result in collision failures.

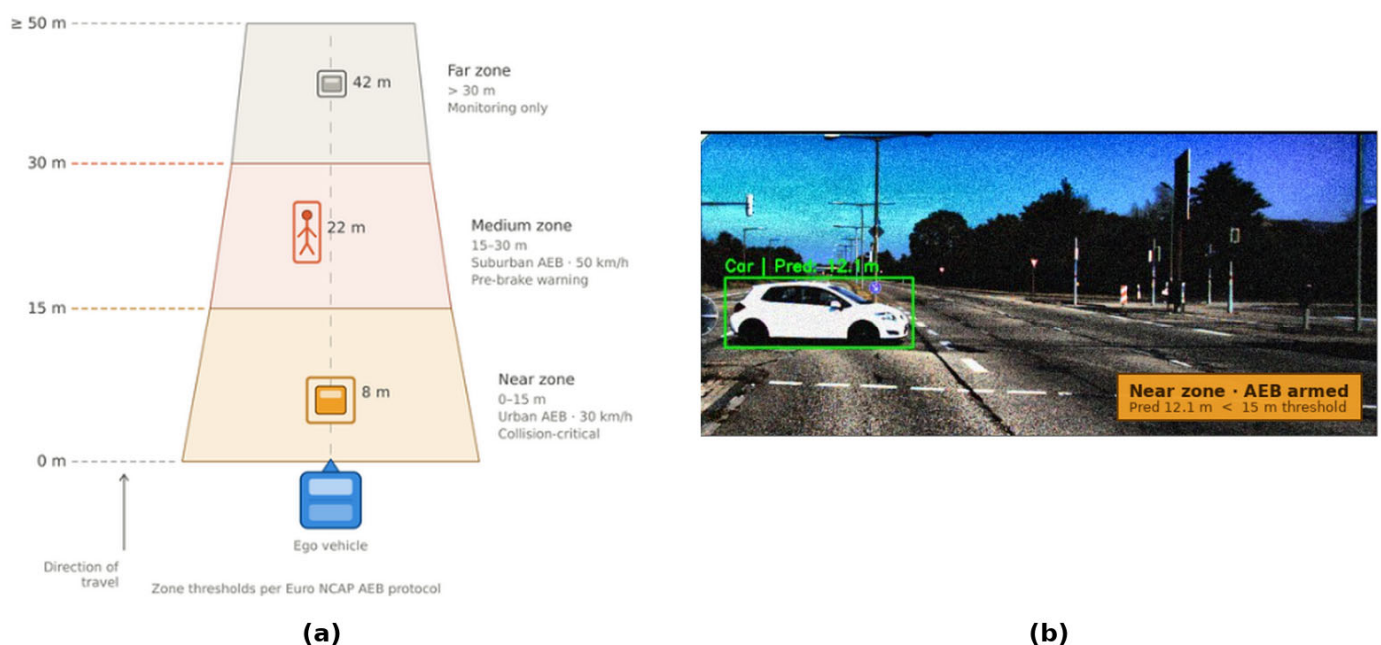


Figure 6. Kinematic safety boundaries of the AGL Autonomous Emergency Braking (AEB) inference engine. (a) Schematic of the three operational zones—Near (0–15 m), Medium (15–30 m), and Far (>30 m)—projected along the ego vehicle’s direction of travel. The 15 m and 30 m thresholds are deliberately aligned with the Euro NCAP AEB stopping-distance requirements for urban (30 km/h) and suburban (50 km/h) operation, respectively [32,33]. Example objects at 8 m, 22 m, and 42 m illustrate how detections are routed to the appropriate kinematic response: collision-critical braking in the Near zone, pre-brake warning in the Medium zone, and passive monitoring in the Far zone. (b) Live frame captured from the AGL inference engine on a Dark-KITTI scene. The vehicle is detected at a predicted distance of 12.1 m, below the 15 m Euro NCAP threshold and triggering the Near-zone AEB response, as indicated by the amber classification badge. This panel demonstrates the end-to-end translation of a 2D bounding box detection into a kinematic safety decision through the pinhole projection defined in Equation (2).

To analyze the localization performance of 2D bounding boxes, we partitioned the target-domain evaluation into three spatial ranges, as shown in Table 2. All AGL results are reported at ($\lambda_y = 2.0$, $\lambda_h = 5.0$) unless otherwise stated; sensitivity to this parameter is analyzed in Section 5.4.

All models were evaluated on the intersection of ground-truth detections by all three models ($\text{IoU} \geq 0.5$), eliminating differential detector recall as a confound.

In the Near range (0–15 m), all three models exhibit comparable performance because the high pixel-to-meter resolution of close-range targets reduces sensitivity to bottom-edge drift, making the anisotropic penalty largely redundant in this regime. The largest performance gap is observed in the Medium driving range (15–30 m), where AGL matches Vanilla YOLOv26n (5.96 m) and modestly outperforms Vanilla YOLOv10n (6.12 m) in the intersection-set comparison. In the Far zone (>30 m), AGL maintains localization accuracy comparable to the baseline methods, reflecting the inherent sensitivity of static pinhole projection at long distances rather than a model-specific limitation. This operating regime corresponds directly to the Euro NCAP 50 km/h stopping-distance boundary, where monocular localization primarily supports monitoring-only functions. In the Far range (>30 m), both models exhibit hyperbolic error scaling consistent with the inverse-square relationship in Equation (3). At this distance, a single-pixel error at the horizon line equates to several meters of depth distortion, highlighting the fundamental physical limit of static monocular focal constraints. We further note that all RMSE values reported in Table 2 are computed under the flat-ground assumption (Section 3.1) with a fixed optical horizon = 165.0; performance on graded terrain (vehicle pitch greater than approximately 3°) is expected to degrade and is reserved for the dynamic pose-compensation extension discussed in Section 7.

While absolute RMSE escalates monotonically with distance (2.42 ± 0.03 m Near, 5.94 ± 0.19 m Medium, 17.41 ± 1.25 m Far), the depth-normalized Abs Rel metric reveals a non-monotonic pattern: 0.264 ± 0.008 (Near), 0.229 ± 0.003 (Medium), 0.311 ± 0.006 (Far). The Medium zone achieves the lowest Abs Rel because it sits in the geometric sweet spot of monocular projection—far enough that absolute pixel errors do not dominate the small ground-truth distances, but not yet entering the inverse-square sensitivity amplification regime where small pixel errors produce multi-meter depth errors. The Near zone Abs Rel (0.264 ± 0.008) is elevated because the denominator Z_{gt} is small, so even sub-meter residuals translate into 25%+ relative error. The Far zone Abs Rel (0.311 ± 0.006) reflects the genuine onset of inverse-square sensitivity beyond 30 m, where the framework approaches the practical geometric limitation of static monocular projection. This decomposition reframes the apparent Far-zone limitation: the framework is not geometrically degrading but instead reflects the fundamental physical limit of static monocular depth estimation at long range. While the principal performance gains are achieved in the Medium zone (15–30 m), the proposed AGL framework maintains Far-zone localization accuracy comparable to existing techniques, including similar δ_1 performance. This behavior supports the Euro NCAP monitoring-only classification beyond 30 m.

5.3. Ablation Study: Resolution Capacity Limits

A central design question for any lightweight monocular framework is whether reducing the input resolution sacrifices accuracy or merely lowers computational cost. We address this question via two complementary ablations: a within-domain resolution sweep on Dark-KITTI (Table 3) and a zero-shot evaluation on real low-light imagery from the ExDark dataset (Figure 7).



Figure 7. Zero-shot generalization performance on the ExDark dataset. (**Top row**): Vanilla YOLOv10n trained on Dark-KITTI. (**Bottom row**): AGL with CLAHE preprocessing. (**Car panel**) Vanilla detects no cars; AGL recovers two cars at the curb. (**Pedestrian panel**) Vanilla misses the pedestrian against the textured wall; AGL recovers with high confidence. (**Cyclist panel**) Vanilla misclassifies the cyclist as ‘Pedestrian’ with high confidence (0.66)—a particularly hazardous failure mode for AEB systems where vehicle-class influences kinematic prediction. AGL correctly classifies the target as ‘Cyclist’. All cases demonstrate cross-domain generalization without ExDark-specific training.

5.3.1. Resolution Sweep on Dark-KITTI

Table 3 reports the accuracy and computational metrics of five configurations spanning two input resolutions (1280 px and 640 px) and three architectural variants: Vanilla YOLOv10n, AGL, and AGL + CLAHE.

Three findings emerge. First, AGL operating at 640 px input resolution achieves RMSE essentially indistinguishable from AGL at 1280 px (10.91 vs. 10.92 m, a 0.01 m difference well within bootstrap uncertainty) while delivering $1.65\times$ the inference throughput (182 vs. 110 FPS) and 33% lower memory consumption (87 vs. 131 MB). The $L_{y_{\text{tall}}}$ geometric prior compensates for reduced spatial sampling, preserving efficient-compute viability without sacrificing accuracy. Second, AGL at 640 px outperforms the Vanilla YOLOv10n baseline at 1280 px across all metrics (RMSE 10.91 vs. 12.14 m; MAE 7.15 vs. 8.10 m), demonstrating that geometric supervision is a more effective strategy for monocular localization than brute-force resolution increase. Third, the full nocturnal configuration AGL + CLAHE at 640 px achieves the lowest RMSE in the study (10.55 m), confirming that the CLAHE preprocessor and the geometric prior contribute complementary improvements rather than competing for the same supervision signal.

5.3.2. Zero-Shot Generalization on ExDark

To assess whether the CLAHE preprocessor and AGL framework generalize beyond the KITTI calibration domain, we evaluated both zero-shot on the ExDark dataset, which contains real low-light imagery from uncalibrated camera perspectives. Figure 7 illustrates a representative case in which the unenhanced baseline fails to detect a pedestrian whose silhouette is submerged in the road’s luminance floor. At the same time, the CLAHE-enhanced inference recovers the same target. The cross-domain transfer indicates that AGL’s lightweight enhancement preprocessing improves nocturnal perception across differ-

ent lighting conditions. Its low computational overhead also suggests potential suitability for future edge-device deployment, although this was not evaluated in the present study.

5.4. Sensitivity Analysis of the Anisotropic Weight

The $L_{y_{\text{tall}}}$ loss introduces two hyperparameters: λ_y , which controls the weight of the positional (bottom-edge) term, and λ_h , which controls the weight of the height-regularization term. Their selection requires care for two reasons:

1. The two coefficients interact non-additively in the loss landscape, so independent single-axis optimization can identify misleading local optima, and
2. configurations at the apparent loss minimum may be initialization-sensitive, producing single-seed results that fail to reproduce. This subsection presents the analyses that informed our selection of $(\lambda_y = 2.0, \lambda_h = 5.0)$ as the production configuration.

5.4.1. Single-Axis Sweeps

Tables 4 and 5 report the depth-estimation RMSE on Dark-KITTI under one-dimensional sweeps of each coefficient with the other held fixed.

Table 4. λ_y sensitivity analysis on Dark-KITTI with $\lambda_h = 5.0$ fixed. Pinhole RMSE reported in meters. **Bold** = minimum in column. The apparent single-axis minimum at $\lambda_y = 1.0$ (Overall RMSE 10.30 m) is not the joint optimum; see Table 5 and Figure 8.

λ_y	Near RMSE	Medium RMSE	Far RMSE	Overall RMSE	N
0.5	2.40 ± 0.03	5.92 ± 0.19	17.34 ± 1.17	10.90 ± 0.70	4500
1.0	2.39 ± 0.03	5.89 ± 0.22	16.39 ± 1.11	10.30 ± 0.59	4482
2.0 (selected)	2.43 ± 0.03	5.94 ± 0.18	17.43 ± 1.24	10.91 ± 0.68	4461
3.0	2.41 ± 0.03	5.74 ± 0.13	16.57 ± 1.02	10.51 ± 0.58	4515
5.0	2.41 ± 0.03	5.84 ± 0.16	17.04 ± 1.18	10.76 ± 0.67	4508

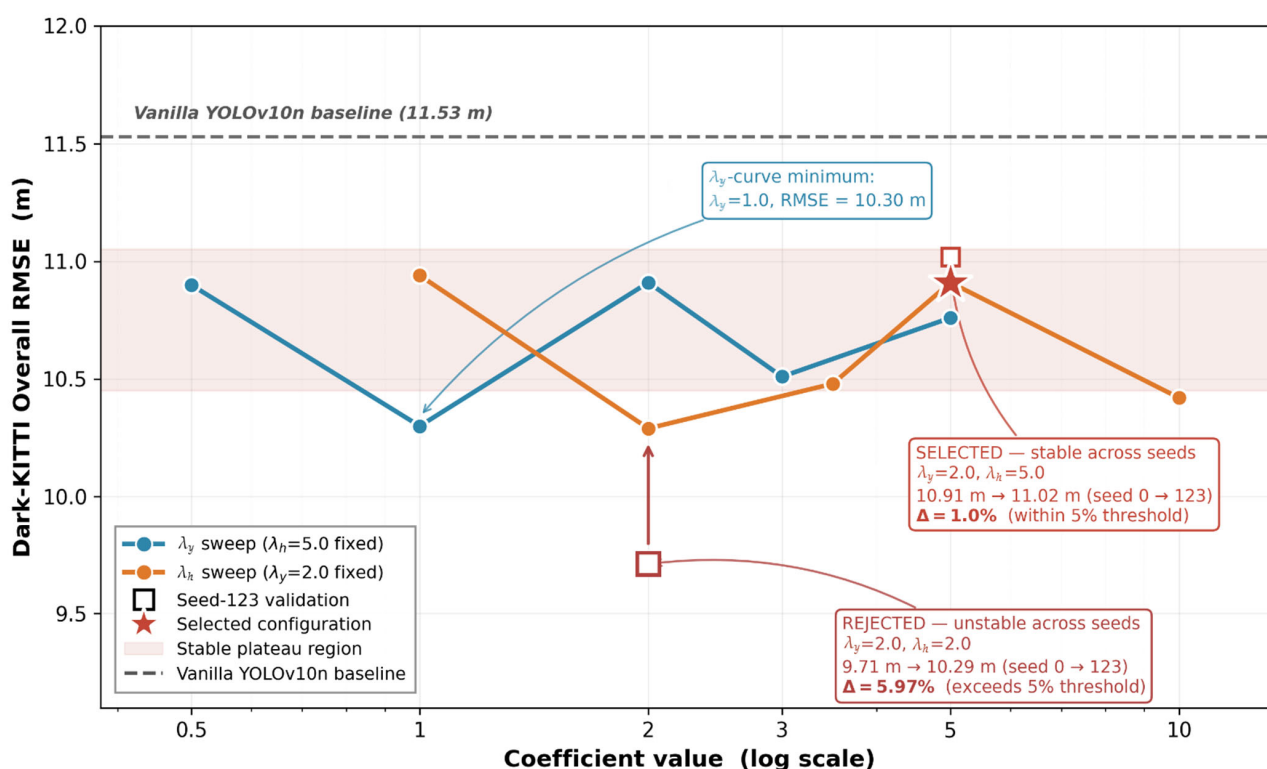


Figure 8. Single-axis sensitivity of the AGL coefficients on Dark-KITTI Overall RMSE. The blue curve sweeps λ_y with λ_h fixed at 5.0 (Table 4); the orange curve sweeps λ_h with λ_y fixed at 2.0 (Table 5).

The two curves attain their minima at different coefficient values ($\lambda_y = 1.0$ at RMSE 10.30 m; $\lambda_h = 2.0$ at RMSE 10.29 m), demonstrating that independent single-axis optimization misidentifies the joint optimum. The apparent global minimum at ($\lambda_y = 2.0$, $\lambda_h = 2.0$) is unstable across initialization seeds (seed 0: 9.71 m $\rightarrow$ seed 123: 10.29 m, $\Delta = 5.97\%$) and is therefore rejected. The selected configuration ($\lambda_y = 2.0$, $\lambda_h = 5.0$, ★) is the cross-seed-stable operating point (seed 0: 10.91 m $\rightarrow$ seed 123: 11.02 m, $\Delta = 1.0\%$).

Table 5. λ_h sensitivity analysis on Dark-KITTI with $\lambda_y = 2.0$ fixed. Pinhole RMSE reported in meters. **Bold** = minimum in column.

λ_h	Near RMSE	Medium RMSE	Far RMSE	Overall RMSE	N
1.0	2.41 $\pm$ 0.03	5.90 $\pm$ 0.16	17.19 $\pm$ 1.11	10.94 $\pm$ 0.63	4667
2.0 *	2.42 $\pm$ 0.03	5.96 $\pm$ 0.20	15.07 $\pm$ 0.28	9.71 $\pm$ 0.18	4624
3.5	2.43 $\pm$ 0.03	5.98 $\pm$ 0.21	16.46 $\pm$ 1.00	10.48 $\pm$ 0.56	4535
5.0 (selected)	2.43 $\pm$ 0.03	5.94 $\pm$ 0.18	17.43 $\pm$ 1.24	10.91 $\pm$ 0.68	4461
10.0	2.41 $\pm$ 0.03	5.94 $\pm$ 0.20	16.56 $\pm$ 1.06	10.42 $\pm$ 0.60	4384

Note: The operating point ($\lambda_y = 2.0$, $\lambda_h = 2.0$ *) with the seed = 0 result (9.71 m) was confirmed as an initialization artifact: under seed = 123, the same configuration yields 10.29 m ($\Delta = 5.97\%$). The selected operating point ($\lambda_y = 2.0$, $\lambda_h = 5.0$) is cross-seed-stable (10.91 $\rightarrow$ 11.02 m, $\Delta = 1.0\%$). See Section 5.4.2.

Read independently, the two sweeps suggest $\lambda_y = 1.0$ (Overall RMSE 10.30 m) and $\lambda_h = 2.0$ (Overall RMSE 9.71 m at seed 0) as individual optima. These per-axis minima, however, do not identify the true joint optimum, and the $\lambda_h = 2.0$ result in particular conceals a seed-stability problem analyzed below.

5.4.2. Joint Sensitivity and Seed-Stability Validation

Figure 8 plots the joint sensitivity landscape, overlays single-axis sweeps, and marks the configurations that received independent seed validation.

The seed-validation outcome at ($\lambda_y = 2.0$, $\lambda_h = 2.0$) is critical to the selection logic. The apparent single-seed minimum (9.71 m at seed 0) shifts to 10.29 m under seed 123, a 5.97% relative variation that exceeds the conventional 5% stability threshold for production deployment. At ($\lambda_y = 2.0$, $\lambda_h = 5.0$), the same dual-seed validation yields 10.91 $\rightarrow$ 11.02 m, a 1.0% variation that lies within ordinary YOLO-class training noise. The selected configuration is therefore the cross-seed-stable operating point, not the single-seed minimum.

The mechanism behind this asymmetry is consistent with a known pattern in multi-task loss optimization. When an auxiliary loss term is weakly weighted relative to the dominant task losses (here, the standard YOLO IoU and classification losses), the optimization trajectory's exploitation of the auxiliary signal becomes initialization-dependent. The bootstrap analysis of Table 5 makes this asymmetry quantitative: at ($\lambda_y = 2.0$, $\lambda_h = 2.0$), the within-seed bootstrap standard deviation on Overall RMSE is 0.18 m, while the between-seed variation reaches 0.58 m—a $3\times$ ratio indicating that the apparent stability of the seed-0 result is a property of the specific initialization rather than of the loss landscape. At ($\lambda_y = 2.0$, $\lambda_h = 5.0$), the height regularization is strongly weighted enough to integrate the geometric prior regardless of initialization, producing seed-invariant behavior across both seeds and a between-seed-comparable bootstrap standard deviation of 0.68 m.

5.4.3. Selected Configuration

Combining the joint sensitivity evidence with the seed-stability validation, we retain ($\lambda_y = 2.0$, $\lambda_h = 5.0$) as the production configuration. This is the only point in our search space that (i) lies in a stable basin of the joint loss landscape, (ii) shows seed-invariant behavior across the validated seeds, and (iii) simultaneously improves both KITTI-Bright (Day) and Dark-KITTI (Night) RMSE over the Vanilla YOLOv10n baseline. Single-seed minima

at ($\lambda_y = 1.0$) and ($\lambda_h = 2.0$) are rejected on the methodological principle that production deployments cannot rely on a favorable initialization.

6. Limitations

While the AGL framework achieves consistent cross-illumination depth estimation at lightweight parameter counts, several limitations bound the scope of the present validation and motivate the directions discussed in Section 7.

6.1. Flat-Ground Assumption

The pinhole projection in Equation (2) assumes a static optical horizon y_{horizon} , equivalent to zero camera pitch. At the KITTI calibration values, a pitch deviation of 1° introduces a horizon shift of ~ 12.6 pixels and a relative depth error of approximately 31% at $Z = 30$ m. KITTI val sequences are predominantly on level urban roads, validating the assumption for the reported numbers, but graded terrain or dynamic vehicle pitch (braking, suspension travel) will exceed this tolerance. A lightweight dynamic pose-estimation head is proposed in Section 7.

6.2. Synthetic Nocturnal Evaluation

Dark-KITTI is algorithmically degraded from KITTI-Bright via gamma + Poisson-Gaussian noise + glare. While this preserves exact 3D ground truth, it does not capture all phenomena of real nocturnal capture (e.g., motion blur from low shutter speeds, sensor-specific noise floors, varied headlight beam patterns). Real-image qualitative corroboration is provided via ExDark (Section 5.3), but quantitative validation on real nocturnal benchmarks (Nighttime Driving Dataset, BDD100K night split) is reserved for future work.

6.3. Wet-Surface CLAHE Behavior

On rain-slicked nocturnal surfaces with specular headlight reflections, CLAHE histogram redistribution may amplify the specular highlight rather than the useful tire-road micro-contrast, potentially injecting noise into the y_{tall} gradient. Quantitative evaluation requires a dedicated wet-surface nocturnal benchmark.

6.4. Quantitative Reporting Scope

RMSE/MAE/Abs Rel/ δ_1 are computed on the Car class only, which provides the most numerous and best-calibrated 3D ground-truth instances. Cross-class generalization across the Pedestrian and Cyclist categories is demonstrated qualitatively on ExDark (Figure 7) but not quantitatively reported.

6.5. Calibration Dependence

The framework depends on per-vehicle camera intrinsics ($f_y, H_{\text{cam}}, y_{\text{horizon}}$). For KITTI, these are fixed at (721.53 px, 1.65 m, 165 px). Deployment on a different sensor requires a one-time per-vehicle recalibration; the framework itself does not need retraining unless the new sensor's image distribution differs substantially from the KITTI training distribution.

6.6. Computational Budget Scope

Full multi-seed training ($N \geq 3$ seeds per evaluated configuration) would exceed the computing budget available for this study. Seed sensitivity is therefore characterized via targeted dual-seed validation on the apparent global minimum and the selected operating point (Figure 8, Section 5.4), with bootstrap confidence intervals on tabulated quantities (Section 4.6) capturing finite-sample uncertainty separately from training stochasticity. Validation on embedded hardware platforms remains future work.

7. Conclusions and Future Work

This work presented AGL (Anisotropic Geometry Loss), a framework that addresses the severe localization degradation of heavyweight 3D detection models under low-light conditions. By replacing dense feature extraction with a geometry-enforcing anisotropic penalty (y_{tall}) and a luminance-channel CLAHE preprocessor, AGL achieves stable cross-domain depth estimation at >160 FPS on a 2.71 M-parameter backbone. The framework demonstrates that explicit geometric constraints provide an effective approach for maintaining robustness under multi-illumination domain shifts, more efficient than scaling either network capacity or input resolution.

A key empirical finding of this work is the non-separability of the joint (λ_y, λ_h) loss landscape. Single-axis sensitivity sweeps yield distinct marginal optima, but the per-axis optima at $\lambda_y = 1.0$ and $\lambda_h = 2.0$ do not jointly identify the global minimum, as confirmed by the seed-instability documented at $(\lambda_y = 2.0, \lambda_h = 2.0)$ in Section 5.4. This result supports theoretically motivated coefficient selection over greedy single-axis tuning and provides a methodological guideline for future anisotropic-loss design. We acknowledge a limitation: training was conducted for 100 epochs to maintain a tractable computational budget across all sweep configurations. Extended training schedules (e.g., 300 epochs) may yield further improvements but are unlikely to alter the qualitative ranking of configurations reported here.

Three directions are reserved for future work. First, a scale-aware dynamic penalty in which λ_y is modulated inversely with the bounding-box height would tighten the geometric constraint specifically on distant targets (>30 m), where pinhole sensitivity is highest, without sacrificing edge-compute performance. Second, the current flat-ground assumption restricts AGL to level topographies; we plan a lightweight pose-estimation head (a single-output regression branch with under 50 K additional parameters) trained on synthetically pitched KITTI raw sequences to dynamically update y_{horizon} per frame, with minimal latency cost, preserving the >160 FPS target. Third, extending the y_{tall} prior to attention-augmented backbones such as YOLOv26 represents a promising direction for combining geometric anchoring with modern feature aggregation.

These results demonstrate that the AGL framework provides a robust, scalable, illumination-invariant solution for AEB-relevant depth estimation on resource-constrained desktop GPUs and provides a promising foundation for future embedded deployment.

Author Contributions: Conceptualization, R.C.; methodology, R.C.; software, R.C.; validation, R.C. and S.-G.M.; formal analysis, R.C.; investigation, R.C.; resources, S.-G.M.; data curation, R.C.; writing—original draft preparation, R.C.; writing—review and editing, R.C. and S.-G.M.; visualization, R.C.; supervision, S.-G.M.; project administration, S.-G.M.; funding acquisition, S.-G.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The KITTI dataset is publicly available at <https://www.cvlibs.net/datasets/kitti/> (accessed on 31 March 2026) and the ExDark dataset at <https://github.com/cs-chan/Exclusively-Dark-Image-Dataset> (accessed on 26 March 2026). The Dark-KITTI synthesis pipeline (gamma + Poisson-Gaussian noise + overexposure mask), the customized Ultralytics v8.4.32 containing the $L_{y_{\text{tall}}}$ regression head and the safety-clamp logic, and the trained model weights at $(\lambda_y = 2.0, \lambda_h = 5.0)$ will be released in a public GitHub repository upon acceptance. Additional experimental results supporting the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors acknowledge the computational resources provided by Chung Yuan Christian University. During the preparation of this manuscript, artificial intelligence (AI)-assisted language tools were used to support drafting, editing, manuscript organization, and formatting review. The authors thoroughly reviewed and revised all AI-assisted outputs and accept full responsibility for the final content of the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

Abbreviation	Meaning
AEB	Autonomous Emergency Braking
AGL	Anisotropic Geometry Loss (proposed framework)
AV	Autonomous Vehicle
CIoU	Complete Intersection over Union
CLAHE	Contrast Limited Adaptive Histogram Equalization
CNN	Convolutional Neural Network
DETR	Detection Transformer
DFL	Distribution Focal Loss
Euro NCAP	European New Car Assessment Programme
FPS	Frames Per Second
GAN	Generative Adversarial Network
GFLOPs	Giga Floating-point Operations per second
IoU	Intersection over Union
KITTI	Karlsruhe Institute of Technology and Toyota Technological Institute (benchmark)
LAB	CIE $L^*a^*b^*$ color space
LiDAR	Light Detection and Ranging
MAE	Mean Absolute Error
NMS	Non-Maximum Suppression
RGB	Red-Green-Blue color space
RMSE	Root Mean Square Error
YOLO	You Only Look Once (object-detector family)

References

1. Romera, E.; Bergasa, L.M.; Yang, K.; Alvarez, J.M.; Barea, R. Bridging the day and night domain gap for semantic segmentation. In Proceedings of the 2019 IEEE Intelligent Vehicles Symposium (IV), Paris, France, 9–12 June 2019; pp. 1312–1318.
2. Sakaridis, C.; Dai, D.; Van Gool, L. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7374–7383.
3. Weng, S.E.; Miaou, S.G.; Christanto, R.; Hsu, C.P. Exposure correction in driving scenes using the atmospheric scattering model. In Proceedings of the 2024 International Conference on Consumer Electronics–Taiwan (ICCE-Taiwan), Taichung, Taiwan, 9–11 July 2024; pp. 493–494. [\[CrossRef\]](#)
4. Brazil, G.; Liu, X. M3D-RPN: Monocular 3D region proposal network for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9287–9296.
5. Lu, Y.; Ma, X.; Yang, L.; Zhang, T.; Liu, Y.; Chu, Q.; Yan, J.; Ouyang, W. Geometry uncertainty projection network for monocular 3D object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 3111–3121.
6. Liu, Z.; Wu, Z.; Tóth, R. SMOKE: Single-stage monocular 3D object detection via keypoint estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020; pp. 996–997.
7. Chen, Y.; Tai, L.; Sun, K.; Li, M. MonoPair: Monocular 3D object detection using pairwise spatial relationships. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 12093–12102.

8. Liu, X.; Xue, N.; Wu, T. Learning auxiliary monocular contexts helps monocular 3D object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022; Volume 36, pp. 1810–1818.
9. Zhang, Y.; Lu, J.; Zhou, J. Objects are different: Flexible monocular 3D object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Online, 19–25 June 2021; pp. 3289–3298.
10. Reading, C.; Harakeh, A.; Chae, J.; Waslander, S.L. Categorical depth distribution network for monocular 3D object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Online, 19–25 June 2021; pp. 8555–8564.
11. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
12. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
13. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-time end-to-end object detection. In Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 10–15 December 2024; Volume 37, pp. 107984–108011. [[CrossRef](#)]
14. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000.
15. Chabot, F.; Chaouch, M.; Rabarisoa, J.; Teulière, C.; Chateau, T. Deep MANTA: A coarse-to-fine many-task network for joint 2D and 3D vehicle analysis from monocular image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2040–2049.
16. Zuiderveld, K. Contrast limited adaptive histogram equalization. In *Graphics Gems IV*; Heckbert, P.S., Ed.; Academic Press Professional: San Diego, CA, USA, 1994; pp. 474–485.
17. Neumann, L.; Karg, M.; Zhang, S.; Scharfenberger, C.; Piegert, E.; Mistr, S.; Prokofyeva, O.; Thiel, R.; Vedaldi, A.; Zisserman, A.; et al. NightOwls: A pedestrians at night dataset. In *Computer Vision—ACCV 2018*; Springer: Cham, Switzerland, 2019; Volume 11361, pp. 691–705. [[CrossRef](#)]
18. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
19. Huang, Z.; Ali, M.A.H.; Nukman, Y.; Xu, H.L.; Zhang, S.; Chen, H.; Alkhedher, M. A Systematic Review of Monocular Depth Estimation for Autonomous Driving: Methods and Dataset Benchmarking. *Results Eng.* **2025**, *26*, 105359. [[CrossRef](#)]
20. Mousavian, A.; Anguelov, D.; Flynn, J.; Kosecka, J. 3D bounding box estimation using deep learning and geometry. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 7074–7082.
21. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.
22. Zhang, C.; Lee, D. Advancing nighttime object detection through image enhancement and domain adaptation. *Appl. Sci.* **2024**, *14*, 8109. [[CrossRef](#)]
23. Sapkota, R.; Cheppally, R.H.; Sharda, A.; Karkee, M. YOLOv26: Key architectural enhancements and performance benchmarking for real-time object detection. *arXiv* **2025**, arXiv:2509.25164.
24. Rezaatofghi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.
25. Sui, X.; Gao, S.; Xu, A.; Zhang, C.; Wang, C.; Shi, Z. Lightweight monocular depth estimation using a fusion-improved transformer. *Sci. Rep.* **2024**, *14*, 22472. [[CrossRef](#)] [[PubMed](#)]
26. Xia, Z.; Wu, T.; Wang, Z.; Zhou, M.; Wu, B.; Chan, C.Y.; Kong, L.B. Dense monocular depth estimation for stereoscopic vision based on pyramid transformer and multi-scale feature fusion. *Sci. Rep.* **2024**, *14*, 7037. [[CrossRef](#)] [[PubMed](#)]
27. Weng, S.E.; Hsu, C.P.; Hsiao, C.Y.; Christanto, R.; Miaou, S.G. From dim to glow: Dynamic illuminance adjustment for simultaneous exposure correction and low-light image enhancement. *Signal Image Video Process.* **2024**, *18*, 8937–8947. [[CrossRef](#)]
28. Wang, Y.; Li, B.; Wei, Z.; Rahman, A.; Gummadi, D.; Zhao, H.; Donzella, V. A survey and new perspective of sensing in the dark for intelligent transportation systems. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 18283–18303. [[CrossRef](#)]
29. Chen, Z.; Li, Z.; Zhang, S.; Fang, L.; Jiang, Q.; Zhao, F.; Zhou, B.; Zhao, H. AutoAlign: Pixel-instance feature aggregation for multi-modal 3D object detection. In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, Vienna, Austria, 23–29 July 2022; pp. 827–833. [[CrossRef](#)] [[PubMed](#)]
30. Weng, X.; Kitani, K. Monocular 3D object detection with pseudo-LiDAR point cloud. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 857–866.
31. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2223–2232.

32. Ku, J.; Pon, A.D.; Waslander, S.L. Monocular 3D object detection leveraging accurate proposals and shape reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 11867–11876.
33. Zhang, C.; Lee, D. EASD: Exposure Aware Single-Step Diffusion Framework for Monocular Depth Estimation in Autonomous Vehicles. *Appl. Sci.* **2025**, *15*, 9130. [[CrossRef](#)]
34. Xiong, M.; Zhang, Z.; Liu, J.; Zhang, T.; Xiong, H. Monocular Depth Estimation Using Self-Supervised Learning with More Effective Geometric Constraints. *Eng. Appl. Artif. Intell.* **2024**, *128*, 107489. [[CrossRef](#)]
35. Weng, S.E.; Miaou, S.G.; Christanto, R. A Lightweight low-light image enhancement network via channel prior and gamma correction. *Int. J. Pattern Recognit. Artif. Intell.* **2025**, *39*, 2554013. [[CrossRef](#)]
36. Wu, Y.; Chen, Y.; Yuan, L.; Liu, Z.; Wang, L.; Li, H.; Fu, Y. Rethinking Classification and Localization for Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10186–10195.
37. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO series in 2021. *arXiv* **2021**, arXiv:2107.08430.
38. Feng, C.; Zhong, Y.; Gao, Y.; Scott, M.R.; Huang, W. TOOD: Task-aligned one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Online, 11–17 October 2021; pp. 3490–3499.
39. Loh, Y.P.; Chan, C.S. Getting to know low-light images with the exclusively dark dataset. *Comput. Vis. Image Underst.* **2019**, *178*, 30–42. [[CrossRef](#)]
40. Eigen, D.; Puhrsch, C.; Fergus, R. Depth map prediction from a single image using a multi-scale deep network. In Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 8–13 December 2014; Volume 27, pp. 2366–2374.
41. Euro NCAP. *Assessment Protocol—Safety Assist: Collision Avoidance, Version 10.4.1*; European New Car Assessment Programme: Brussels, Belgium, 2024. Available online: https://cdn.euroncap.com/cars/assets/euro_ncap_assessment_protocol_sa_collision_avoidance_v1041_fafe1dd418.pdf (accessed on 24 May 2026).
42. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; Chapman & Hall/CRC: New York, NY, USA, 1993.
43. Mohamed, S.A.E.; Alshalfan, K.A.; Al-Hagery, M.A.; Othman, M.T.B. Safe driving distance and speed for collision avoidance in connected vehicles. *Sensors* **2022**, *22*, 7051. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.