symmetry

MDPI

Article

RGC-Mono: Ratio-Guided Geometry Calibration for Monocular 3D Object Detection

Yuhang Zhao [1] and Ao Xu [2,*]

[1] College of Mechanic and Control Engineering, Guilin University of Technology, Guilin 541004, China; 18860368161@163.com
[2] Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University, Shanghai 200092, China
* Correspondence: aoxu@tongji.edu.cn

Abstract

Monocular 3D detection estimates metric depth from ambiguous single-view geometry. RGC-Mono calibrates a perspective-derived depth anchor with a dimensionless log-ratio residual. Scale-Normalized Projection Ratio Residual (SPRR) represents multiplicative projection discrepancy, Geometry Sensitivity-Aware Uncertainty (GSAU) weights depth regression during training, and Orientation-Visibility Conditioned Calibration (OVCC) uses pose and visibility cues to condition the correction. With a shared pretrained MonoDGP checkpoint and 250 additional training epochs, three KITTI Car validation runs obtain Moderate $AP_{3D}|_{R40}$ of 22.91 ± 0.14% for matched MonoDGP, 23.26 ± 0.12% for ordinal-OVCC RGC-Mono, and 23.36 ± 0.08% for categorical OVCC. The respective mean gains are 0.35 and 0.45 percentage points. A capacity-matched SPRR–additive comparison gives a mean difference of 0.10 points with a confidence interval spanning zero, locating the detection benefit in the integrated configuration. The original ordinal checkpoint reaches 19.49% Moderate AP3D on KITTI test under its reported protocol. The framework provides a compact way to combine relative correction, geometric uncertainty, and visibility context for car depth localization.

Keywords: monocular 3D object detection; autonomous driving; geometric depth; ratio calibration; uncertainty estimation; Transformer

1. Introduction

Reliable three-dimensional perception is fundamental to autonomous driving because planning and control require accurate estimates of surrounding objects, free space, and collision risk. LiDAR and stereo systems provide strong range measurements at the cost of additional sensing, calibration, and deployment complexity. Monocular 3D object detection offers a compact camera-only alternative by estimating object category, 3D position, dimensions, and orientation from a single RGB image. The missing range measurement makes depth estimation the central challenge: perspective projection compresses different 3D configurations into similar image observations, leaving object localization weakly constrained. Recent monocular detectors therefore introduce geometric structure, depth-sensitive supervision, and uncertainty modeling to strengthen depth reasoning [1–7].

Research in monocular 3D detection has evolved from direct attribute regression toward increasingly structured depth representations. Geometry-aware methods exploit projection relations between image measurements and physical object dimensions, auxiliary-task methods enrich image features with depth-sensitive supervision, and multi-depth methods

Check for updates

Academic Editor: Jie Yang

Received: 10 October 2026
Revised: 3 November 2026
Accepted: 10 November 2026
Published: 17 November 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) license.

combine estimates derived from complementary cues. Transformer-based detectors further aggregate scene-level geometry through depth-aware object queries [1–10]. Recent approaches extend this progression through geometric proposal refinement, complementary depth hypotheses, occupancy learning, and foreground-depth supervision [11–14]. These developments consistently show that the representation used for depth learning has a direct influence on 3D localization accuracy.

MonoDGP advances geometry-based depth reasoning by computing a perspective-derived geometric depth from the predicted 2D box height and 3D object height, followed by an additive geometry-error correction [15]. Its error target is substantially more concentrated than raw metric depth and provides a strong baseline for image-only monocular detection. The additive correction remains expressed in meters. For a fixed projection-discrepancy factor, the idealized projection relation makes the additive target proportional to metric depth. Small projected heights and noisy size predictions further affect the geometric anchor in practical images.

The central problem is thus the representation of geometry correction. An effective correction variable should preserve the interpretable perspective anchor while describing projection discrepancy in a form that is comparable across depth ranges. Geometry-related uncertainty can modulate depth-regression gradients through the training objective. At the same time, the visual envelope defining a 2D box varies systematically with orientation, occlusion, and truncation, which changes the correction required to recover central depth.

Research Question. *Can the integration of SPRR, training-time GSAU, and OVCC improve KITTI Car depth localization over a checkpoint- and training-budget-matched MonoDGP baseline?* We assess this question using parameter-matched correction, component, and occlusion-stratified ablations.

We introduce **RGC-Mono**, a Ratio-Guided Geometry Calibration framework built around this representation problem. **Scale-Normalized Projection Ratio Residual** (**SPRR**) replaces metric additive correction with a logarithmic depth ratio and reconstructs central depth multiplicatively. **Geometry Sensitivity-Aware Uncertainty** (**GSAU**) models geometry-related predictive variance for heteroscedastic weighting of the depth loss during training; the resulting variance weights the log-depth loss. **Orientation-Visibility Conditioned Calibration** (**OVCC**) conditions the ratio prediction on orientation, occlusion, and truncation cues to model systematic changes in the visible object envelope. The resulting pipeline follows a single progression from geometric anchoring, through reliability estimation, to relative depth calibration. Recent studies on camera-height robustness, test-time geometric uncertainty, intrinsic variation, and spatial-projection consistency further emphasize the importance of reliable geometric modeling for monocular 3D detection [16–19].

In summary, our contributions are threefold:

1. We derive the depth-scale factor in additive geometry correction and introduce SPRR, which represents the correction as a log ratio relative to a perspective-derived anchor.
2. GSAU combines log-height and residual variances for geometry-aware loss weighting during training. OVCC conditions the correction mean on orientation and visibility.
3. We evaluate the integrated framework on KITTI Car using a shared initialization and equal additional training budgets. Three-run Moderate validation $AP_{3D|R40}$ is 22.91 ± 0.14% for MonoDGP, 23.26 ± 0.12% for ordinal OVCC, and 23.36 ± 0.08% for categorical OVCC, giving mean gains of 0.35 and 0.45 percentage points. Capacity-matched correction heads and component ablations characterize the contribution of each design.

2. Related Work

Extra training data. The lack of direct depth cues makes monocular 3D detection inherently underconstrained, motivating the use of additional supervision or transferable depth information. DD3D pre-trains dense depth from large-scale monocular data and transfers the learned geometry to 3D detection [20], while CaDDN predicts categorical pixel-depth distributions and lifts image features into a 3D representation [21]. LiDAR-guided approaches further exploit point clouds to construct stronger monocular supervision or pseudo labels [22], and multi-view consistency has been used to improve semi-supervised monocular detectors [23]. Beyond in-domain supervision, STMono3D studies cross-domain adaptation under camera and depth distribution shifts [24], OPA-3D explicitly aggregates visible pixels for occluded objects [25], and MonoTTA adapts a pretrained detector to unlabeled test distributions [26]. More recent work reduces dependence on expensive 3D annotation through scalable pseudo supervision [27] or sparse-label learning [28], while open-set formulations extend monocular detection to unseen scenes and categories [29]. These methods demonstrate the value of richer supervision and adaptation across different training and generalization settings. RGC-Mono follows the standard image-only detection protocol and studies the representation used to calibrate an explicit geometric depth anchor.

Transformer-based methods. CNN-based monocular detectors established strong image-space baselines and aggregate long-range geometric context progressively through stacked local operations. MonoDTR introduces depth-aware Transformer features to couple visual appearance with spatial reasoning [3], while MonoDETR adopts object queries and a depth-guided decoder for end-to-end 3D prediction [8]. MonoUNI extends this idea toward unified vehicle- and infrastructure-side detection [9], and MonoPGC injects pixel-level geometry contexts into monocular features [10]. Subsequent methods improve query construction or geometric verification, including 3D bounding-box denoising with a Perceiver-style refinement stage [11] and adaptive token allocation in MonoATT [30]. Other Transformer or query-based frameworks move toward richer intermediate representations: MonoNeRD employs NeRF-like geometry [31], UniMODE constructs a unified BEV detector across heterogeneous scenes [32], and MonoDiff models 3D attributes with a diffusion process [33]. Recent methods further explore open-set geometry-aware queries [29], ranked retrieval of multiple plausible hypotheses [34], and conditional chains of 3D attribute prediction [35]. These architectures improve contextual reasoning and representation capacity. RGC-Mono builds on the established Transformer detector and concentrates on the depth variable optimized by its geometry branch, allowing the effect of geometry-correction parameterization to be studied directly.

Multi-depth estimation modes. Early monocular methods directly regressed metric depth together with other 3D attributes, using pairwise relations, temporal motion, truncation-aware formulations, feature alignment, or localization-oriented training to stabilize prediction [36–40]. Geometry-based methods subsequently introduced physical projection constraints to reduce the ambiguity of direct regression. GUPNet models uncertainty in geometric projection [1], Geometry-Based Distance Decomposition factors distance through object and image geometry [41], MonoDDE constructs diverse depth solutions from multiple geometric cues [4], and densely constrained or probabilistic formulations further increase the number and reliability of available depth hypotheses [42,43]. Motion and ground-plane constraints provide additional geometric conditions [7,44], while DID-M3D decomposes instance depth into visual and attribute components [5], DEVIANT learns depth-equivariant representations [6], and MonoEdge exploits local perspective cues [45]. More recent work emphasizes complementarity among multiple depths [12], robustness to camera height or test-time uncertainty [16,17], and sensitivity to intrinsic parameters or 3D-to-2D projection consistency [18,19]. MonoDGP uses perspective-derived geometric

depth as an anchor and predicts an additive geometry error as a compact alternative to multi-depth branches [15]. Its error distribution is substantially more compact than raw depth, while the additive residual remains coupled to metric-depth scale. RGC-Mono follows the same geometry-error perspective while changing the correction from an additive metric residual to a scale-normalized log ratio. This preserves the interpretable projection prior and isolates the remaining discrepancy in a relative form, which is then calibrated according to geometric sensitivity, orientation, and visibility.

Recent evaluation settings. CHARM3R [16] studies generalization to unseen camera mounting heights using extended CARLA scenes, and MonoSAOD [28] evaluates sparse-label KITTI learning. These settings complement the fully annotated, standard-camera KITTI protocol used here. SPAN [18], RARE [34], and MonoCoP [35] report standard KITTI Car results and enter the numerical comparison. MonoUNI [9] includes a vehicle-side image-only KITTI experiment within its broader vehicle–infrastructure formulation.

Logarithmic and relative depth representations. Eigen et al. [46] introduced a scale-invariant objective based on log-depth differences, and DORN [47] used spacing-increasing discretization for ordinal regression. ZoeDepth [48] combines relative-depth representations with metric adaptation, while ScaleDepth [49] decomposes metric depth into scene scale and relative depth. RGC-Mono applies log-ratio regression at the object level: predicted physical height, projected box height, and focal length define a perspective anchor, and the network learns the correction relative to that anchor.

Heteroscedastic uncertainty and geometric propagation. Kendall and Gal [50] developed input-dependent aleatoric regression objectives, while Kendall et al. [51] used task-level homoscedastic uncertainties for multi-task weighting. GUPNet [1] and GUPNet++ [52] incorporate geometric-projection uncertainty into monocular detection. Building on these principles, GSAU combines log-height variance components with a log-ratio correction variance. Its perspective-derived anchor, multiplicative correction, and diagonal log-depth variance model operate within MonoDGP's geometry-error formulation. Sections 3.3 and 4 examine the associated covariance approximation.

Table 1 summarizes the relationships between RGC-Mono and prior depth representations and uncertainty formulations.

Table 1. Relationship to prior depth representations and uncertainty formulations.

Representative Approach	Depth Variable/Objective	Geometric Anchor	Uncertainty Treatment
Eigen et al. [46]; DORN [47]	Scale-invariant log differences/ordinal depth bins	Scene/pixel depth formulation	–
ZoeDepth [48]; ScaleDepth [49]	Relative depth with metric scale recovery	Scene-level metric scaling	–
Kendall and Gal [50]; Kendall et al. [51]	General regression/multi-task learning	Task-dependent	Input-dependent heteroscedastic/task-level homoscedastic weighting
GUPNet [1]; GUPNet++ [52]	Geometry-projected metric depth	Object-size projection	Geometry-aware uncertainty propagation
MonoDGP [15]	Additive metric-depth correction	Object-size projection	Baseline depth uncertainty
RGC-Mono	**Anchored log-ratio correction**	**Object-size projection**	**Log-height and ratio-variance decomposition**

3. Method

3.1. Overview

The overall framework of **RGC-Mono** follows the decoupled visual-depth Transformer architecture of MonoDGP [15], while replacing its additive geometry-error branch with a ratio-domain calibration mechanism. Given a monocular image $I \in \mathbb{R}^{H_I \times W_I \times 3}$, a ResNet-50 backbone extracts multi-scale features $\{f_i\}_{i=0}^{3}$. The Region Segmentation Head (RSH) retained from MonoDGP enhances foreground responses and provides segment embeddings for the depth branch. The enhanced visual features f_V and depth features f_D are then encoded by the parallel visual and depth encoders. A 2D visual decoder first produces object queries q_{2d} for category, 2D bounding box, and projected-center prediction. These queries initialize the depth-guided 3D decoder, which generates q_{3d} for object dimensions, orientation, and depth reasoning. This organization keeps the original 2D/3D query decoupling intact and localizes our changes to the geometry-calibration stage, following the overall design philosophy of MonoDGP. The reference method similarly separates visual and depth branches and applies its geometry correction after the geometric depth is computed. Figure 1 summarizes the calibration pathway. The network predicts the mean correction μ_r, with training target $r_{\text{target}} = \log Z^{\text{gt}} - \log \hat{Z}_{\text{geo}}$.

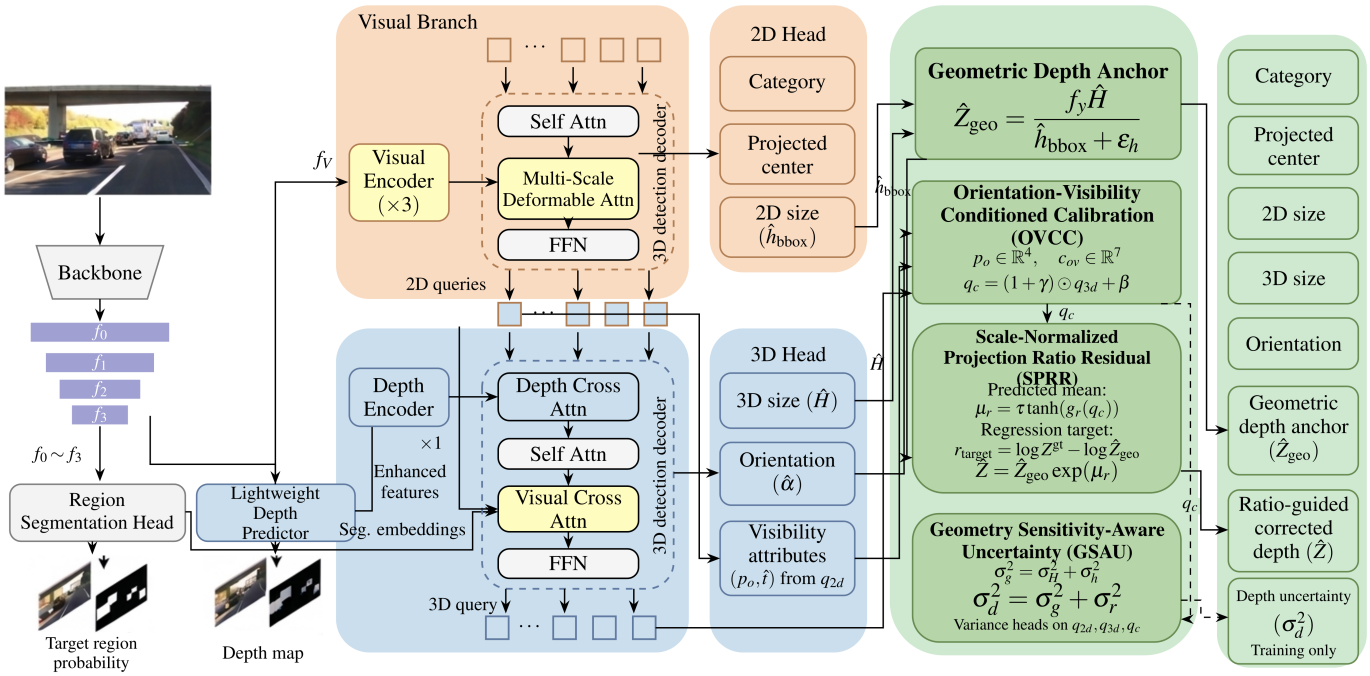


Figure 1. RGC-Mono calibration pathway. Shared visual-depth features provide geometric attributes and object queries. Categorical OVCC conditions SPRR's mean correction; GSAU constructs the variance used in the training loss. Inference depth follows $\hat{Z} = \hat{Z}_{\text{geo}} \exp(\mu_r)$.

KITTI axial-depth and object-center convention. We adopt the rectified KITTI reference-camera coordinate system, in which X points to the right, Y points downward, and Z points forward. Throughout the paper, $Z > 0$ is the *camera-fram axial depth* in meters, equivalently the third coordinate of the KITTI 3D location, and is not the Euclidean range $R = \sqrt{X^2 + Y^2 + Z^2}$. The KITTI 3D location $B = (X_b, Y_b, Z_b)^\top$ specifies the *bottom-face center* of the object box. For physical box height H, its geometric center is $C = (X_b, Y_b - H/2, Z_b)^\top$; the two centers share the same axial depth Z_b. The projected center (u_c, v_c) predicted by the 2D branch denotes the perspective projection of C, not the midpoint of the axis-aligned 2D detection box. The box height h_{bbox} in Equation (1) is instead the pixel height of that

2D image envelope. These distinctions keep the supervision of object-center depth, image center, and projected box height consistent with KITTI annotations [53].

For each matched object, the 2D head predicts a box height $\hat{h}_{\text{bbox}}$, while the 3D head predicts the physical object height $\hat{H}$. With the known vertical focal length f_y, we first compute the geometric depth anchor

$$\hat{Z}_{\text{geo}} = \frac{f_y \hat{H}}{\hat{h}_{\text{bbox}} + \varepsilon_h} \tag{1}$$

Here, f_y and $\hat{h}_{\text{bbox}}$ are measured in pixels, while $\hat{H}$ and $\hat{Z}_{\text{geo}}$ are measured in meters. The height-domain stabilization constant is fixed at $\varepsilon_h = 10^{-3}$ pixels. For a positive projected height h, denote by $Z_0 = f_y H/h$ the geometric depth without stabilization and by $Z_\varepsilon = f_y H/(h + \varepsilon_h)$ its stabilized counterpart. The exact relative perturbation satisfies

$$\frac{|Z_\varepsilon - Z_0|}{Z_0} = \frac{\varepsilon_h}{h + \varepsilon_h} < \frac{\varepsilon_h}{h},$$

while the absolute log-depth perturbation is bounded by

$$\left|\log Z_\varepsilon - \log Z_0\right| = \log(1 + \varepsilon_h/h) \leq \varepsilon_h/h.$$

The KITTI 3D detection protocol uses minimum ground-truth 2D box heights of 40 pixels for Easy and 25 pixels for Moderate/Hard [53]. The corresponding perturbations are 0.0025% and 0.0040%, respectively. For a 5-pixel-high box, the analytical perturbation is 0.0200%. These bounds describe numerical stabilization; prediction errors in box height contribute separately to geometric-depth error.

RGC-Mono predicts a log-ratio correction μ_r from the geometric anchor and reconstructs the central depth multiplicatively:

$$\hat{Z} = \hat{Z}_{\text{geo}} \exp(\mu_r). \tag{2}$$

Calibration source and geometric back-projection. For each KITTI left-color image, the calibrated projection matrix $P_2 \in \mathbb{R}^{3\times4}$ is read from its associated calibration file [53]. In the rectified reference-camera coordinates used by the 3D labels, its pinhole form is

$$P_2 = \begin{bmatrix} f_x & 0 & c_x & t_x \\ 0 & f_y & c_y & t_y \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

Thus $f_x = P_2(1,1)$ and $f_y = P_2(2,2)$ are focal lengths in pixels, and $c_x = P_2(1,3)$ and $c_y = P_2(2,3)$ are principal-point coordinates in pixels (one-based matrix indexing). The fourth-column entries $t_x = P_2(1,4)$ and $t_y = P_2(2,4)$ have units of pixel $\cdot$ meter and are projection-matrix offsets. The projected geometric center obeys

$$\hat{u}_c = \frac{f_x \hat{X}_c + c_x \hat{Z} + t_x}{\hat{Z}}, \qquad \hat{v}_c = \frac{f_y \hat{Y}_c + c_y \hat{Z} + t_y}{\hat{Z}}.$$

The corrected axial depth and the predicted projected 3D center recover the geometric center in the rectified camera frame through

$$\begin{aligned} \hat{X}_c &= \frac{(\hat{u}_c - c_x)\hat{Z} - t_x}{f_x}, \\ \hat{Y}_c &= \frac{(\hat{v}_c - c_y)\hat{Z} - t_y}{f_y}. \end{aligned} \tag{3}$$

The KITTI bottom-face center is then $(\hat{X}_b, \hat{Y}_b, \hat{Z}_b) = (\hat{X}_c, \hat{Y}_c + \hat{H}/2, \hat{Z})$. Substitution into P_2 yields the original projected point $(\hat{u}_c, \hat{v}_c)$, establishing the algebraic consistency of projection and back-projection. When $t_x = t_y = 0$, or when the reference frame is transformed so that these offsets are absorbed consistently, Equation (3) reduces to the usual intrinsic-only form. Pixel coordinates and the projection matrix always refer to the same image coordinate system: for a 2D affine resize/crop transformation A, transformed image coordinates use $P_2' = AP_2$; equivalently, mapped-back original pixel coordinates use the original P_2. The predicted $\hat{Z}$ remains axial depth.

Three components are introduced around Equation (2). First, **Scale-Normalized Projection Ratio Residual (SPRR)** reparameterizes geometry correction in the log-ratio domain. Second, **Geometry Sensitivity-Aware Uncertainty (GSAU)** estimates the reliability of the projection anchor from the uncertainties of the quantities used in Equation (1). Third, **Orientation-Visibility Conditioned Calibration (OVCC)** explains the remaining object-dependent ratio variation using orientation, occlusion, and truncation cues. The three components form a single reasoning chain: SPRR defines the relative correction variable, GSAU estimates the reliability of the geometric anchor, and OVCC models object-dependent variation of that correction.

3.2. Scale-Normalized Projection Ratio Residual

Perspective projection provides a strong prior for monocular depth estimation. Under the standard assumptions used by geometry-based monocular detectors, the ideal projected central height h_c of an object with physical height H and central depth Z satisfies

$$h_c = \frac{f_y H}{Z}. \tag{4}$$

In practice, the 2D bounding-box height follows the visible image envelope, while the ideal central projection is defined by the object-center geometry. We write

$$h_{\text{bbox}} = \kappa h_c, \tag{5}$$

where κ is a dimensionless projection-discrepancy factor. It absorbs the difference between the axis-aligned visual box and the latent central projection. MonoDGP also attributes the difference between h_{bbox} and h_c to the visual surface of the object and uses this discrepancy to motivate geometry-error correction.

If the matched object height is exact, substituting Equation (5) into the geometric projection yields

$$Z_{\text{geo}} = \frac{f_y H}{h_{\text{bbox}}} = \frac{Z}{\kappa}. \tag{6}$$

For an additive geometry-error formulation,

$$e_Z = Z - Z_{\text{geo}} = Z\left(1 - \frac{1}{\kappa}\right). \tag{7}$$

For exact physical height and fixed κ, Equation (7) makes the metric correction proportional to Z. This scaling relation motivates a relative correction variable.

RGC-Mono represents geometry correction by the logarithmic depth ratio

$$r = \log \frac{Z}{Z_{\text{geo}}}. \tag{8}$$

Using Equation (6),

$$r = \log \kappa = \log \frac{h_{\text{bbox}}}{h_c}. \tag{9}$$

For fixed κ, $r = \log \kappa$ is independent of the explicit metric-depth factor. With predicted dimensions and boxes, the empirical target also reflects height, pose, visibility, and calibration errors. The controlled correction-head comparison in Section 4 evaluates the detection behavior of this parameterization.

Relation to direct log-depth regression. A direct log-depth predictor has the form $\log \hat{Z} = g_\theta(q)$. SPRR instead uses the explicit decomposition

$$\log \hat{Z} = \log \hat{Z}_{\text{geo}} + \mu_r = \log f_y + \log \hat{H} - \log(\hat{h}_{\text{bbox}} + \varepsilon_h) + \mu_r.$$

The perspective-derived anchor, bounded residual head, and residual supervision specify the model. Under Equations (4)–(6), equal multiplicative projection discrepancies yield the same target at different metric depths. Section 4.3 compares additive, linear-ratio, and log-ratio correction within the detector.

The 3D query is mapped to an unbounded scalar a_r, and the mean ratio correction is parameterized as

$$\mu_r = \tau \tanh(a_r), \qquad a_r = g_r(q_c), \tag{10}$$

where $g_r(\cdot)$ is a two-layer MLP, q_c is the calibrated query defined in Section 3.4, and τ bounds the correction range for numerical stability. When OVCC is disabled in the ablation study, we set $q_c = q_{3d}$, allowing SPRR and GSAU to be evaluated independently of visibility conditioning. The final depth is obtained by Equation (2). Because $\exp(\mu_r) > 0$, the multiplicative formulation produces positive depth values, while the bounded tanh parameterization limits excessively large corrections.

The natural regression error for SPRR is measured in log depth,

$$\Delta_r = \log Z_{\text{gt}} - \log \hat{Z}_{\text{geo}} - \mu_r. \tag{11}$$

At the optimum, μ_r estimates the effective relative correction required by the current geometric anchor. Since the residual is defined in the same logarithmic domain as the projection ratio, its magnitude is directly comparable across depth ranges.

Structural coupling of the regression target. Define a reference anchor, $Z^*_{\text{geo}} = f_y H_{\text{gt}} / (h^{\text{gt}}_{\text{bbox}} + \varepsilon_h)$, anchor error $e_g = \log \hat{Z}_{\text{geo}} - \log Z^*_{\text{geo}}$, and reference correction $r^* = \log Z_{\text{gt}} - \log Z^*_{\text{geo}}$. The effective training target satisfies

$$r_{\text{target}} = \log Z_{\text{gt}} - \log \hat{Z}_{\text{geo}} = r^* - e_g, \qquad \text{Cov}(e_g, r_{\text{target}} - r^*) = -\text{Var}(e_g).$$

The anchor error therefore enters the target deviation with a negative sign. The learned correction error $e_r = \mu_r - r^*$ also depends on the network prediction and shared features. Equation (11) gives $\Delta_r = -(e_g + e_r)$, linking the residual loss to final log-depth error.

3.3. Geometry Sensitivity-Aware Uncertainty

The reliability of the scale-normalized correction depends on the quality of the geometric anchor. Equation (1) uses both the predicted 3D height and the predicted 2D box height. A small perturbation in either quantity changes the geometric depth, and this effect becomes especially pronounced when a distant object occupies only a few image pixels.

Taking the logarithm of Equation (1) gives

$$\log \hat{Z}_{\text{geo}} = \log f_y + \log \hat{H} - \log(\hat{h}_{\text{bbox}} + \varepsilon_h) \tag{12}$$

We model the uncertainties of $\log \hat{H}$ and $\log \hat{h}_{\text{bbox}}$, which places both terms on the same relative scale as the ratio correction. Two lightweight uncertainty heads predict

$$\sigma^2_H = \text{Softplus}(g_H(q_{3d})) + \varepsilon_v, \qquad \sigma^2_h = \text{Softplus}(g_h(q_{2d})) + \varepsilon_v \tag{13}$$

where σ_H^2 and σ_h^2 denote variances in the log-height domain. The variance-domain stabilization constant is $\varepsilon_v = 10^{-6}$ (dimensionless), distinct from the pixel-valued ε_h in Equation (1).

Under a diagonal-covariance approximation for the log-height errors, the geometric-anchor variance becomes:

$$\sigma_g^2 = \sigma_H^2 + \sigma_h^2. \tag{14}$$

Let $e_H = \log \hat{H} - \log H_{\text{gt}}$ and $e_h = \log(\hat{h}_{\text{bbox}} + \varepsilon_h) - \log(h_{\text{bbox}}^{\text{gt}} + \varepsilon_h)$. The geometric-anchor error $e_g = e_H - e_h$ satisfies

$$\text{Var}(e_g) = \text{Var}(e_H) + \text{Var}(e_h) - 2\,\text{Cov}(e_H, e_h).$$

Equation (14) uses the diagonal approximation by omitting the cross-covariance. Section 4 quantifies this omitted contribution across depth ranges.

This formulation is consistent with first-order relative-error propagation. In particular, for small perturbations,

$$d \log Z_{\text{geo}} = \frac{dH}{H} - \frac{dh_{\text{bbox}}}{h_{\text{bbox}} + \varepsilon_h} \tag{15}$$

which explains why a similar absolute pixel error leads to larger relative uncertainty when h_{bbox} is small.

More precisely, the first-order sensitivity to a log-height perturbation is $h_{\text{bbox}}/(h_{\text{bbox}} + \varepsilon_h)$. Equation (14) approximates its square by one; at the default ε_h, the resulting coefficient differs from one by less than 0.04% even at a box height of 5 pixels. The stated variance addition therefore retains the same first-order accuracy in the evaluated height regime.

Heteroscedastic log-domain likelihoods supervise σ_H^2 and σ_h^2, so that each variance tracks the prediction error of its corresponding geometric quantity:

$$\mathcal{L}_H^u = \frac{(\log \hat{H} - \log H_{\text{gt}})^2}{2\sigma_H^2} + \frac{1}{2} \log \sigma_H^2, \tag{16}$$

$$\mathcal{L}_h^u = \frac{(\log \hat{h}_{\text{bbox}} - \log h_{\text{bbox}}^{\text{gt}})^2}{2\sigma_h^2} + \frac{1}{2} \log \sigma_h^2. \tag{17}$$

The ratio branch predicts a residual variance

$$\sigma_r^2 = \text{Softplus}(g_\sigma(q_c)) + \varepsilon_v \tag{18}$$

The ratio branch contributes a non-negative residual term to the log-depth variance proxy. The diagonal model combines the anchor and correction contributions as

$$\sigma_d^2 = \sigma_g^2 + \sigma_r^2. \tag{19}$$

The RGC depth objective is then written as

$$\mathcal{L}_{\text{RGC}} = \frac{\Delta_r^2}{2\sigma_d^2} + \frac{1}{2} \log \sigma_d^2. \tag{20}$$

Heteroscedastic depth objective. Equations (16), (17), and (20) follow Gaussian negative log-likelihood regression with input-dependent variances [50]. GSAU constructs the depth variance from separately supervised log-height components and a residual component, extending geometric-projection uncertainty principles [1,52] to the calibrated log-depth variable. The task-level weighting formulation in [51] uses homoscedastic uncertainties.

The second variance addition also omits anchor–correction covariance. With $e_r = \mu_r - r^*$ and final log-depth error $e_d = e_g + e_r$, the empirical identity is

$$\text{Var}(e_d) = \text{Var}(e_g) + \text{Var}(e_r) + 2\,\text{Cov}(e_g, e_r).$$

The experiment reports $\mathrm{Corr}(e_g, e_r)$ to characterize this dependence. Empirical component errors and per-instance predicted variances describe different statistical quantities.

Variance allocation. Writing $V = \sigma_g^2 + \sigma_r^2$, Equation (20) depends on the sum V. For any δ preserving positivity,

$$(\sigma_g^2, \sigma_r^2) \mapsto (\sigma_g^2 + \delta, \sigma_r^2 - \delta)$$

leaves the depth likelihood unchanged, and

$$\frac{\partial \mathcal{L}_{\mathrm{RGC}}}{\partial \sigma_g^2} = \frac{\partial \mathcal{L}_{\mathrm{RGC}}}{\partial \sigma_r^2} = \frac{V - \Delta_r^2}{2V^2}.$$

The auxiliary height objectives in Equations (16) and (17) constrain the geometric contribution. The components are jointly learned contributions to a diagonal variance proxy; their separate identifiability requires additional constraints.

Use during training and inference. The three predicted variances enter the auxiliary objectives and Equation (20). Holding other predictions fixed,

$$\frac{\partial \mathcal{L}_{\mathrm{RGC}}}{\partial \mu_r} = -\frac{\Delta_r}{\sigma_d^2}.$$

The variance thus weights the depth-regression gradient during training. Inference reconstructs depth using $\hat{Z} = \hat{Z}_{\mathrm{geo}} \exp(\mu_r)$ and retains the baseline classification-based postprocessing. GSAU's ablation jointly evaluates its variance heads, auxiliary supervision, and loss weighting.

3.4. Orientation-Visibility Conditioned Calibration

The scale-normalized residual varies across objects because κ depends on which visible surfaces determine the 2D box envelope. Object orientation changes the projected contour, while occlusion and truncation may remove the upper or lower extremal pixels that define h_{bbox}. MonoDGP similarly observes that geometry error is affected by object attributes and orientation. We condition the ratio prediction on these factors so that the calibration head can model this structured residual variation.

The orientation $\hat{\alpha}$ is obtained from the original 3D orientation head. We further attach a lightweight visibility head to q_{2d}. It predicts a four-category occlusion distribution

$$p_o = \mathrm{Softmax}(g_o(q_{2d})) \in \mathbb{R}^4, \tag{21}$$

and a normalized truncation ratio

$$\hat{t} = \mathrm{Sigmoid}(g_t(q_{2d})). \tag{22}$$

KITTI labels 0, 1, 2, and 3 denote fully visible, partly occluded, largely occluded, and unknown states [53]. Categories 0–2 describe qualitative severity, while category 3 is unspecified. OVCC preserves this distinction through a categorical probability embedding:

$$\mathbf{e}_o = \sum_{k=0}^{3} p_o^{(k)} \mathbf{e}_k = [p_o^{(0)}, p_o^{(1)}, p_o^{(2)}, p_o^{(3)}]^\top \in \mathbb{R}^4 \tag{23}$$

where $\mathbf{e}_k \in \mathbb{R}^4$ is the one-hot basis vector for category k. The orientation–visibility descriptor is

$$c_{ov} = [\sin\hat{\alpha},\ \cos\hat{\alpha},\ \mathbf{e}_o^\top,\ 1 - \hat{t}] \in \mathbb{R}^7. \tag{24}$$

Sine and cosine encode angular periodicity, the four probabilities preserve occlusion-category identity, and $1 - \hat{t}$ encodes truncation. The modulation MLP accepts seven features;

the original ordinal implementation used four. The categorical cross-entropy and truncation objectives in Equation (27) supervise both variants.

The descriptor modulates the 3D query through a feature-wise affine transformation. Specifically,

$$[\gamma, \beta] = g_{ov}(\mathrm{sg}(c_{ov})), \tag{25}$$

$$q_c = (1 + \gamma) \odot q_{3d} + \beta, \tag{26}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. The stop-gradient operation keeps orientation and visibility governed by their dedicated supervision and decouples these semantic predictions from ratio-loss optimization. The calibrated query q_c is then used by Equations (10) and (18) to predict the ratio mean and residual uncertainty.

The visibility branch retains the same categorical supervision as the original implementation:

$$\mathcal{L}_{\mathrm{vis}} = \mathrm{CE}(p_o, o_{\mathrm{gt}}) + \lambda_t \, \mathrm{SmoothL1}(\hat{t}, t_{\mathrm{gt}}). \tag{27}$$

OVCC conditions the SPRR correction while preserving the geometric anchor as the primary depth estimate. Orientation and visibility explain systematic variation in the ratio between the anchor and the central depth within the same calibration pathway.

The categorical probabilities provide semantic context for correction; visible-area fraction is a separate physical quantity.

3.5. Geometric Interpretation of Ratio-Guided Calibration

The motivation of RGC-Mono can be summarized from three aspects.

Additive error vs. ratio error. MonoDGP demonstrates that geometry-error prediction is easier than raw depth regression because the error distribution is more compact. Its analysis compares raw depth, dimension errors, and depth errors to explain why geometry correction is effective. RGC-Mono retains this insight and further distinguishes distribution compactness from scale decoupling. Under the ideal projection relation, the additive target is

$$e_Z = Z\left(1 - \frac{1}{\kappa}\right), \tag{28}$$

whereas the proposed ratio target is

$$r = \log \kappa. \tag{29}$$

For a fixed κ, the additive target changes linearly with metric depth, whereas the ratio target contains no explicit Z factor. RGC-Mono therefore models the relative projection discrepancy directly.

Metric-depth loss vs. log-ratio loss. Scale normalization is also applied in the optimization domain. With fixed uncertainty, Equation (20) reduces to a squared error on Δ_r. Its gradient with respect to the predicted ratio mean is

$$\frac{\partial \mathcal{L}_{\mathrm{RGC}}}{\partial \mu_r} = -\frac{\Delta_r}{\sigma_d^2}, \tag{30}$$

whose magnitude is governed by the log-domain residual and its uncertainty. A metric-depth objective on the multiplicative prediction introduces a factor proportional to $\hat{Z}$ in the gradient. The log-domain objective keeps the optimization consistent with the scale-normalized representation.

Residual variation vs. geometric reliability. The ratio varies with object appearance and observation conditions. Pixel quantization, inaccurate 2D boxes, dimension errors, truncation, occlusion, and orientation all perturb the geometric anchor. RGC-Mono separates these effects according to their role. GSAU models how reliable the projection variables

are, while OVCC models systematic object-dependent changes in the visual envelope. This leads to the decomposition

$$\begin{aligned} \underbrace{\log \hat{Z}}_{\text{predicted central depth}} &= \underbrace{\log \hat{Z}_{\text{geo}}}_{\text{geometry anchor}} + \underbrace{\mu_r}_{\text{relative correction}}, \\ \underbrace{\sigma_d^2}_{\text{depth uncertainty}} &= \underbrace{\sigma_g^2}_{\text{anchor uncertainty}} + \underbrace{\sigma_r^2}_{\text{residual uncertainty}}. \end{aligned} \tag{31}$$

The two equations align the mean and uncertainty decomposition in the same log-depth space, linking SPRR, GSAU, and OVCC within one calibration formulation.

3.6. Loss Function

RGC-Mono preserves the original supervision for region segmentation, object-level depth features, 2D detection, and non-depth 3D attributes. Following the baseline, the 2D loss contains category classification, 2D box regression, GIoU, and projected-center terms, while the 3D attribute loss supervises dimensions and orientation. MonoDGP organizes its training objective around region, depth-map, 2D, and 3D losses. The original central-depth term is replaced by the ratio-domain calibration objective, together with the lightweight uncertainty and visibility supervision described above.

The uncertainty supervision for the geometric variables is

$$\mathcal{L}_{\text{sens}} = \mathcal{L}_H^u + \mathcal{L}_h^u. \tag{32}$$

The final objective is

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{2D} + \mathcal{L}_{3D}^{\text{attr}} + \lambda_d \mathcal{L}_{\text{RGC}} + \lambda_s \mathcal{L}_{\text{sens}} + \lambda_v \mathcal{L}_{\text{vis}} + \lambda_m \mathcal{L}_{\text{dmap}} + \lambda_r \sum_{i=0}^{4} \mathcal{L}_{\text{region}}^i, \tag{33}$$

where $\mathcal{L}_{3D}^{\text{attr}}$ excludes the original central-depth regression term, and $\lambda_d, \lambda_s, \lambda_v, \lambda_m, \lambda_r$ balance the corresponding losses. During inference, the uncertainty and visibility heads add small MLP operations, while the backbone, RSH, encoders, and decoders retain the baseline structure. The detector therefore remains image-only and preserves the computational structure of the MonoDGP baseline while modifying the representation and calibration of geometric depth.

4. Experiments

4.1. Settings

Dataset. We evaluate RGC-Mono on the KITTI 3D object detection benchmark [53], which contains 7481 training images and 7518 test images. Following the commonly used split, the 7481 labeled images are divided into 3712 training images and 3769 validation images. KITTI defines Easy, Moderate, and Hard difficulty levels according to object size, occlusion, and truncation. **Evaluation category.** All detection, depth-error, uncertainty, and ablation results in this study concern KITTI Car. Comparison methods use their reported Car metrics under the corresponding protocols.

Evaluation Metrics. We adopt the official AP_{3D} and AP_{BEV} metrics at 40 recall positions, denoted as $AP_{3D}|_{R40}$ and $AP_{BEV}|_{R40}$. For the Car category, a 3D IoU threshold of 0.7 is used. The Moderate $AP_{3D}|_{R40}$ score is treated as the main metric. In addition to detection accuracy, we report depth error over different distance ranges to examine the scale-normalization hypothesis and use uncertainty-error correlation to evaluate whether GSAU reflects the reliability of geometric depth.

Axial-depth evaluation and binning. All central-depth results use the camera-frame Z coordinate of the object center, which equals Z_b in the KITTI bottom-center 3D location annotation. For N matched detections, the reported depth mean absolute error is

$$\mathrm{MAE}_Z = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{Z}_i - Z_i^{\mathrm{gt}}\right| \quad \text{(meters)}.$$

The depth-stratified analyses use the *ground-truth axial depth* Z_i^{gt}, with bins $0 < Z_i^{\mathrm{gt}} \le 20$ m, $20 < Z_i^{\mathrm{gt}} \le 40$ m, and $Z_i^{\mathrm{gt}} > 40$ m. No 3D Euclidean range $\sqrt{X^2+Y^2+Z^2}$ is used for these bins or the central-depth MAE. The three depth bins are specific to our distance-stratified analysis; KITTI Easy/Moderate/Hard are separate visibility- and box-size-based detection difficulty levels.

Uncertainty is evaluated using Spearman correlation with the absolute log-central-depth error. This measures error ranking; the analysis below defines each model's score and distinguishes it from the empirical covariance diagnostics.

Implementation Details. We use ResNet-50 as the backbone and retain the visual-depth Transformer structure of MonoDGP [15]. Each attention module uses eight heads, with 50 object queries and four sampling points in multi-scale deformable attention. The model is trained on a single NVIDIA RTX 4090 GPU with 24 GB of memory. We use AdamW with an initial learning rate of 2×10^{-4}, a weight decay of 10^{-4}, and a batch size of 8.

Training schedule. The 250-epoch stage is additional fine-tuning from pretrained MonoDGP checkpoint C_0. The learning rate is multiplied by 0.5 at epochs 85, 125, 165, and 225.

Mixed-precision training is enabled to reduce memory consumption. The ratio bound in Equation (10) is set to $\tau = 1.0$. The loss weights in Equation (33) are set to $\lambda_d = 1.0$, $\lambda_s = 0.2$, $\lambda_v = 0.2$, $\lambda_m = 1.0$, and $\lambda_r = 1.0$. During inference, queries with category confidence below 0.2 are removed. Categorical OVCC passes the four softmax probabilities into the seven-dimensional descriptor in Equation (24). Validation ablations report ordinal and categorical variants separately.

Initialization and matched training budget. RGC-Mono inherits the ResNet-50 backbone, RSH, visual-depth encoders, and Transformer decoders from pretrained MonoDGP checkpoint C_0; new calibration heads are randomly initialized. The matched MonoDGP control resumes the unmodified detector from the same C_0 for 250 additional epochs. Both configurations use the same data split, image preparation, augmentation, optimizer, learning-rate schedule, evaluation script, and checkpoint-selection rule. Architecture-specific heads retain their respective configurations. Published MonoDGP scores follow its original 250-epoch training protocol and are listed as literature results.

Repeated training and paired statistics. The matched MonoDGP, SPRR-only, SPRR+GSAU, ordinal-OVCC, and categorical-OVCC configurations each have three independent 250-epoch fine-tuning runs. Shared components start from C_0, and runs are paired across configurations by trial identifier. Moderate validation $AP_{3D|R40}$ is summarized by arithmetic mean and sample standard deviation ($n = 3$, divisor $n-1$). Paired differences use a two-sided Student t test and 95% confidence interval with two degrees of freedom, assuming approximately normal run-level differences. An exact two-sided sign-flip test supplements this analysis. These statistics describe training variability; image-sampling uncertainty would require per-image AP resampling. Literature values and single-checkpoint measurements are labeled separately.

4.2. Main Results

Table 2 and Table 3 report the KITTI Car test and validation comparisons, respectively.

We compare RGC-Mono with representative monocular 3D detectors on the KITTI test and validation sets. The selected methods cover geometry-based depth reasoning,

Transformer-based detection, complementary depth estimation, and geometry-error correction, including GUPNet [1], MonoDTR [3], DID-M3D [5], MonoUNI [9], MonoDETR [8], MonoCD [12], FD3D [14], and MonoDGP [15]. The principal numerical comparison uses published KITTI Car metrics; the table identifies auxiliary supervision and pretrained resources, while the text specifies training-protocol differences.

Recent detection architectures. The comparison includes MonoMAE [54] for occlusion-aware query reconstruction, MonoCoP [35] for chained attribute prediction, SPAN [18] for spatial/projection alignment, RARE [34] for candidate retrieval, MonoIA [19] for intrinsic-aware conditioning, and MonoMH [55] for multi-hypothesis localization. Their published or official KITTI scores provide architectural context.

Table 4 summarizes evaluation settings. CHARM3R and MonoSAOD address camera-height shift and sparse annotation, respectively. MonoUNI contributes its vehicle-side KITTI result. The matched validation control in Table 5 evaluates the calibration framework within a common detector and training budget.

Table 2. KITTI car detection comparison. Test $AP|_{R40}$ (IoU 0.7, %). Extra data denotes auxiliary training resources. Published methods and the original ordinal RGC-Mono checkpoint follow their respective training protocols.

Method	Extra Data	AP_{BEV}			AP_{3D}		
		Easy	Mod.	Hard	Easy	Mod.	Hard
GUPNet [1]	None	-	-	-	20.11	14.20	11.77
MonoDTR [3]	LiDAR	28.59	20.38	17.14	21.99	15.39	12.73
DID-M3D [5]	LiDAR	32.95	22.76	19.83	24.40	16.29	13.75
MonoUNI [9]	None	-	-	-	24.75	16.73	13.49
MonoDETR [8]	None	33.60	22.11	18.60	25.00	16.47	13.58
MonoCD [12]	None	33.41	22.81	19.57	25.53	16.59	14.53
FD3D [14]	None	34.20	23.72	20.76	25.38	17.12	14.50
MonoMAE [54]	None	34.15	24.93	21.76	25.60	18.84	16.78
MonoDGP (published) [15]	None	35.24	25.23	22.02	26.35	18.72	15.97
MonoCoP [35]	None	36.77	25.57	22.62	27.54	19.11	16.33
MonoDGP + SPAN [18]	None	-	-	-	27.02	19.30	16.49
RARE [34]	None	38.46	26.73	23.46	28.83	19.57	17.38
MonoIA [19]	VLM prior	37.55	26.59	23.26	29.52	20.29	17.93
MonoMH [55]	None	37.85	28.06	24.53	29.12	20.88	17.93
RGC-Mono (ordinal)	None	35.88	25.96	22.71	27.18	19.49	16.68

Table 3. KITTI Car detection comparison. Validation $AP|_{R40}$ (IoU 0.7, %). Dashes denote unavailable matching values. RGC-Mono uses original ordinal OVCC; matched continuation and repeated-run analyses are reported below.

Method	Extra Data	AP_{BEV}			AP_{3D}		
		Easy	Mod.	Hard	Easy	Mod.	Hard
GUPNet [1]	None	31.07	22.94	19.75	22.76	16.46	13.72
MonoDTR [3]	LiDAR	33.33	25.35	21.68	24.52	18.57	15.51
DID-M3D [5]	LiDAR	31.10	22.76	19.50	22.98	16.12	14.03
MonoUNI [9]	None	-	-	-	24.51	17.18	14.01
MonoDETR [8]	None	37.86	26.95	22.80	28.84	20.61	16.38
MonoCD [12]	None	34.60	24.96	21.51	26.45	19.37	16.38
FD3D [14]	None	36.98	26.77	23.16	28.22	20.23	17.04
MonoMAE [54]	None	40.26	27.08	23.14	30.29	20.90	17.61
MonoDGP (published) [15]	None	39.40	28.20	24.42	30.76	22.34	19.02
MonoCoP [35]	None	-	-	-	-	-	-
MonoDGP + SPAN [18]	None	-	-	-	30.98	23.26	20.17
RARE [34]	None	-	-	-	-	-	-
MonoIA [19]	VLM prior	-	-	-	33.61	24.40	20.80
MonoMH [55]	None	-	-	-	-	-	-
RGC-Mono (ordinal)	None	40.10	28.95	25.10	31.64	23.31	19.88

Table 4. Evaluation settings of recent methods. Standard KITTI Car results appear in Table 2; camera-height shift and sparse-label settings are discussed separately.

Method	Primary Research/Evaluation Setting	Treatment in the KITTI Accuracy Comparison
CHARM3R [16]	Unseen camera mounting heights; extended CARLA evaluation	Camera-height robustness context
MonoSAOD [28]	KITTI with sparsely annotated 3D labels (e.g., 30/50/70% annotated)	Sparse-label supervision context
MonoUNI [9]	Unified vehicle–infrastructure formulation, including vehicle-side KITTI	KITTI vehicle-side image-only score retained and protocol identified
SPAN [18]; RARE [34]; MonoCoP [35]	Reported standard KITTI Car detection metrics	Quantitative entry in Table 2

Table 5. Initialization and training-budget comparison. KITTI Car validation $AP_{3D|R40}$ (IoU 0.7, %). The last three rows use shared checkpoint C_0 and 250 additional epochs. Values are single-checkpoint scores; three-run summaries are reported in the repeated-training analysis below. Published MonoDGP follows its original training protocol.

Method	Initialization	Additional Epochs	Easy ↑	Moderate ↑	Hard ↑
MonoDGP (published) [15]	Original training (250 epochs)	–	30.76	22.34	19.02
MonoDGP (matched continuation)	Identical C_0	250	31.20	22.90	19.55
RGC-Mono (ordinal OVCC)	Shared weights from C_0	250	31.64	23.31	19.88
RGC-Mono (categorical OVCC)	Shared weights from C_0	250	**31.70**	**23.39**	**19.97**

Matched validation results. The continued-training MonoDGP control obtains 31.20/22.90/19.55% Easy/Moderate/Hard $AP_{3D|R40}$. Ordinal RGC-Mono reaches 31.64/23.31/19.88%, giving gains of 0.44/0.41/0.33 percentage points. Categorical OVCC reaches 31.70/23.39/19.97%, giving gains of 0.50/0.49/0.42 points. These are single-checkpoint comparisons; the corresponding three-run mean Moderate gains are 0.35 and 0.45 points (see the paired statistical analysis below).

The KITTI test result of 19.49% Moderate AP3D belongs to the ordinal-OVCC checkpoint. Categorical OVCC is evaluated on validation. Table 2 provides the published test context, whose training protocols differ from the matched continuation experiment.

Training supervision and published-score provenance. The extra-data column identifies auxiliary supervision or pretrained resources: MonoDTR and DID-M3D use LiDAR-related training supervision, while MonoIA uses pretrained vision-language representations [3,5,19]. All methods take images at inference. MonoIA's official KITTI entry reports 20.29% test Moderate AP3D, whereas its released-checkpoint page lists 19.11%; the comparison uses the official entry. Backbone, pretraining, and optimization differences apply to these literature scores.

Position among recent architectures. Ordinal RGC-Mono's 19.49% test Moderate AP3D is above MonoMAE (18.84%), MonoCoP (19.11%), and MonoDGP + SPAN (19.30%), and below RARE (19.57%), MonoIA (20.29%), and MonoMH (20.88%). These methods explore complementary directions including hypothesis modeling, retrieval, intrinsic conditioning, and spatial alignment. RGC-Mono concentrates on geometric-depth calibration within MonoDGP.

The improvement is also reflected in AP_{BEV}, since the corrected central depth directly affects the recovered X and Z coordinates of the 3D center. The gain in AP_{3D} is the primary result for this study because it jointly reflects depth, orientation, and 3D box localization.

Figure 2a–c shows improvements for occlusion, long range, and truncation. Panel (d) retains a 0.35 m depth error in dense traffic, compared with 2.38 m for MonoDGP. Panels (e) and (f) show failures: RGC-Mono's errors are 1.78 and 1.91 m, exceeding MonoDGP's 0.04 and 0.18 m. The examples show both useful corrections and cases where calibration moves depth away from ground truth.



Figure 2. Qualitative comparison on KITTI validation. (**a**–**c**) show heavy occlusion, long range, and truncation; (**d**) shows residual error in dense traffic. Failure cases (**e**,**f**) show larger errors for RGC-Mono than MonoDGP. Each panel gives the ground-truth depth and the two predicted depths with absolute axial-depth errors. Orange denotes ground truth, blue MonoDGP, and green RGC-Mono; dotted boxes indicate zoom regions. Results use original ordinal OVCC.

4.3. Ablation Study

We conduct ablation studies on the KITTI validation split. The RGC-Mono ablation variants share the same backbone, RSH, visual-depth encoders, decoupled queries, and additional training schedule. The MonoDGP geometry-error formulation is compared both as the published literature result and as the newly reported checkpoint- and budget-matched continued-training control. We progressively replace the additive depth error with SPRR,

introduce GSAU, and finally add OVCC. Following the style of the baseline study, all results are reported under the same evaluation protocol.

Computational cost. The proposed modules are implemented with small MLP heads and element-wise operations. Table 6 reports the measured computational cost on a single RTX 4090 with batch size 1. SPRR introduces the ratio head, while GSAU and OVCC add scalar uncertainty and visibility predictions. The original ordinal-OVCC model remains close to MonoDGP in computational cost; the accuracy reference in that table is published MonoDGP. Table 6 reports three-run mean ± SD for Moderate AP3D of methods (1)–(3). Parameters, FLOPs, and latency describe the original ordinal implementation; latency is a batch-1 checkpoint measurement. Categorical OVCC is evaluated separately for accuracy.

Geometry correction mode. We first isolate the main representation change and compare several depth prediction modes using the same detector. Direct depth regression learns central depth without an explicit projection anchor. Geometric depth provides this anchor and remains sensitive to the discrepancy between the box envelope and the object center. MonoDGP reduces this error with an additive metric residual. SPRR corrects the geometric anchor through the log ratio in Equation (8). The single-checkpoint matched MonoDGP scores 22.90% Moderate AP3D, compared with 22.75% for SPRR-only and 22.34% for published MonoDGP. The integrated-model comparison uses the matched control in Table 5.

Table 7 compares the single-checkpoint geometry-correction modes on KITTI Car validation.

Table 6. Computational cost on RTX 4090 (batch 1). Moderate validation AP3D for methods (1)–(3) is three-run mean ± SD; published MonoDGP is a literature score. Parameters, FLOPs, and checkpoint latency describe the original ordinal implementation.

Method	Val, Mod. AP_{3D}	Params (M) ↓	FLOPs (G) ↓	Runtime (ms) ↓
MonoDGP (published) [15]	22.34	38.90	68.99	31.0
method(1): SPRR	22.79 ± 0.14	39.02	69.05	31.2
method(2): SPRR + GSAU	23.08 ± 0.09	39.18	69.12	31.6
method(3): SPRR + GSAU + OVCC (ordinal)	**23.26 ± 0.12**	39.31	69.20	32.1

Table 7. Geometry-correction modes on KITTI Car validation ($AP_{3D|R40}$, %). All values are single-checkpoint scores. Published MonoDGP and matched continuation are separate references; the capacity-matched MLP control reports Moderate AP3D.

Depth Prediction Mode	Easy AP_{3D}	Mod. AP_{3D}	Hard AP_{3D}
Direct Depth	26.60	19.23	16.47
Geometric Depth ($\hat{h}_{\text{bbox}}$)	9.15	7.99	5.52
GD + Additive Depth Error (published) [15]	30.76	22.34	19.02
GD + Additive Depth Error (matched continuation)	**31.20**	**22.90**	**19.55**
GD + Additive Depth Error (capacity-matched MLP)	–	22.62	–
GD + Linear Relative Ratio	30.91	22.55	19.17
GD + Log Projection Ratio (SPRR)	31.18	22.75	19.36

Capacity-matched correction heads. Additive-Matched uses SPRR's two-layer MLP structure, hidden width, input query, nonlinearities, and scalar output dimension. Both complete detectors have 39.02 M parameters, omit GSAU and OVCC, inherit shared weights from C_0, and receive the same 250-epoch continuation budget. The output is interpreted as metric additive correction in one case and bounded log-ratio correction in the other; target and reconstruction transform vary together. Table 8 reports paired-run Moderate AP3D.

SPRR exceeds Additive-Matched by a mean of 0.103 points (sample SD 0.112), with paired differences $[+0.13, -0.02, +0.20]$. The paired t test gives $p = 0.252$ and a 95% interval of $[-0.176, +0.383]$ points; the exact sign-flip p-value is 0.50. The interval spans zero, and

both isolated heads score below matched MonoDGP on average. The positive matched-baseline gain is observed for the integrated framework.

The logarithmic form also has a geometric interpretation. A linear relative ratio such as $Z/Z_{\text{geo}} - 1$ is dimensionless; its distribution remains asymmetric when the geometric anchor underestimates or overestimates the central depth by reciprocal factors. The logarithm maps reciprocal scale errors to opposite values with equal magnitude, which is consistent with the multiplicative geometry in Equation (2).

Table 8. Capacity-matched correction heads on KITTI validation (Car, Moderate $AP_{3D}|_{R40}$, %). All three-run configurations start from the same shared MonoDGP checkpoint and receive 250 additional epochs; run identifiers are paired across methods. Parameters are given for the complete detector. Values are independently trained runs and the corresponding sample mean ± SD.

Method	Parameters (M)	Run 1	Run 2	Run 3	Mean ± SD
MonoDGP (matched continuation)	38.90	22.90	22.77	23.05	22.91 ± 0.14
Additive-Matched (same MLP as SPRR)	39.02	22.62	22.69	22.75	22.69 ± 0.07
SPRR-only (log projection ratio)	39.02	22.75	22.67	22.95	22.79 ± 0.14

Overall performance. Table 9 progressively introduces the three proposed components. The three-run Moderate means progress from 22.79 ± 0.14% for SPRR to 23.08 ± 0.09% with GSAU, 23.26 ± 0.12% with ordinal OVCC, and 23.36 ± 0.08% with categorical OVCC. Adding GSAU further improves the Moderate and Hard results by accounting for the reliability of $\hat{H}$ and $\hat{h}_{\text{bbox}}$. The original OVCC gives an additional gain by conditioning the ratio correction on orientation and visibility.

Table 9. Main ablation study of RGC-Mono on KITTI validation. Three-run Moderate AP3D. For matched methods, Moderate AP3D is mean ± SD; Easy/Hard are single-checkpoint scores. The published MonoDGP row is a literature result.

Method	SPRR	GSAU	OVCC	Easy AP_{3D}	Mod. AP_{3D}	Hard AP_{3D}
MonoDGP (published) [15]				30.76	22.34	19.02
MonoDGP (matched continuation)				31.20	22.91 ± 0.14	19.55
method(1)	✓			31.18	22.79 ± 0.14	19.36
method(2)	✓	✓		31.42	23.08 ± 0.09	19.63
method(3): ordinal OVCC	✓	✓	✓	31.64	23.26 ± 0.12	19.88
method(4): categorical OVCC	✓	✓	✓	**31.70**	**23.36 ± 0.08**	**19.97**

Occlusion encoding. Table 10 compares OVCC removal, the original ordinal expectation $\hat{o} = \sum_{k=0}^{3}(k/3)p_o^{(k)}$, a known-class signal $s_{\text{known}} = \frac{1}{2}p_o^{(1)} + p_o^{(2)}$ with a separate unknown probability $p_o^{(3)}$, and the complete categorical embedding. The three OVCC descriptors contain 4, 5, and 7 features. The variants share the backbone, geometric losses, training schedule, split, and evaluation protocol.

Table 10. Ablation of occlusion encoding on KITTI validation (Car, $AP_{3D}|_{R40}$, %). These are single-checkpoint scores; repeated-run Moderate results for available configurations are reported in Table 11. The original ordinal and revised category-preserving encodings are reported separately.

Occlusion Encoding	Easy AP_{3D} ↑	Moderate AP_{3D} ↑	Hard AP_{3D} ↑
Without OVCC	31.42	23.02	19.63
Ordinal scalar (original)	31.64	23.31	19.88
Known-level ordinal + unknown indicator	31.67	23.35	19.93
Four-class probability embedding	**31.70**	**23.39**	**19.97**

Categorical OVCC improves the single-checkpoint Easy/Moderate/Hard scores over ordinal OVCC by 0.06/0.08/0.09 points. Across three runs, the Moderate increment is 0.10

points with paired-t interval $[-0.01, 0.21]$. The categorical descriptor is the default formulation in Equations (23) and (24) because it preserves the unknown state's category identity.

Repeated-training results. Table 11 reports the individual trials. The ordinal and categorical models improve matched MonoDGP by 0.35 and 0.45 points on average. These estimates average three trials, while Table 5 reports the corresponding single-checkpoint comparisons.

Table 11. Three-run validation accuracy. KITTI Car Moderate $AP_{3D}|_{R40}$ (%). All methods start from the same shared checkpoint C_0 and use 250 additional training epochs. Runs are paired by the same trial identifiers across methods; SD uses the $n-1$ divisor ($n=3$).

Method	Run 1	Run 2	Run 3	Mean ± SD
MonoDGP (matched continuation)	22.90	22.77	23.05	22.91 ± 0.14
method(1): SPRR	22.75	22.67	22.95	22.79 ± 0.14
method(2): SPRR + GSAU	23.02	23.04	23.18	23.08 ± 0.09
method(3): ordinal OVCC	23.31	23.12	23.35	23.26 ± 0.12
method(4): categorical OVCC	23.39	23.27	23.42	**23.36 ± 0.08**

For within-run differences $d_s = AP_{A,s} - AP_{B,s}$, we report $\bar{d} = \frac{1}{3}\sum_s d_s$, sample standard deviation s_d, and interval $\bar{d} \pm t_{0.975,2} s_d/\sqrt{3}$. The ordinal and categorical full models have intervals $[0.22, 0.49]$ and $[0.27, 0.63]$ points against matched MonoDGP. Adding GSAU to SPRR gives $[0.11, 0.47]$ points. The categorical-versus-ordinal interval spans zero (Table 12).

Table 12. Paired statistical comparisons. KITTI validation Moderate AP3D. Differences and confidence intervals are in percentage points. The t-test p-values are two-sided with two degrees of freedom; exact two-sided sign-flip $p = 0.25$ for all displayed three-pair comparisons (all observed differences are positive).

Comparison (***A*–*B***)	Mean **Δ** ± SD (pp)	Paired ***t*** 95% CI (pp)	Paired ***t*****: *p***	Exact Sign-Flip: ***p***
Ordinal OVCC–matched MonoDGP	+0.35 ± 0.06	[+0.22, +0.49]	0.0080	0.25
Categorical OVCC–matched MonoDGP	+0.45 ± 0.07	[+0.27, +0.63]	0.0084	0.25
SPRR + GSAU–SPRR	+0.29 ± 0.07	[+0.11, +0.47]	0.0200	0.25
Categorical OVCC–ordinal OVCC	+0.10 ± 0.04	[−0.01, +0.21]	0.0579	0.25

The paired-t intervals assume approximately normal run-level differences. With three positive pairs, the exact two-sided sign-flip test gives $p = 0.25$, reflecting the limited resolution of three-run inference. Easy/Hard AP3D, depth MAE, and test scores remain single-checkpoint measurements; the paired tests concern the repeated Moderate validation metric.

Scale-normalization and depth-range analysis. Figure 3a reports mean absolute additive residual $|e_Z|$ in meters with image-level 95% bootstrap intervals. Panel (b) compares the dimensionless normalized additive residual

$$\eta = \frac{|e_Z|}{\hat{Z}_{\text{geo}}}$$

with $|r|$. Objectwise, $\eta = |\exp(r) - 1|$, approaching $|r|$ for small corrections. Separate axes preserve the distinction between metric and normalized quantities.

We further divide validation objects according to their ground-truth central depth and report the depth MAE in Table 13. The improvement is modest for near objects and becomes more visible as distance increases, which is consistent with the increasing sensitivity of small projected boxes described by Equation (15).

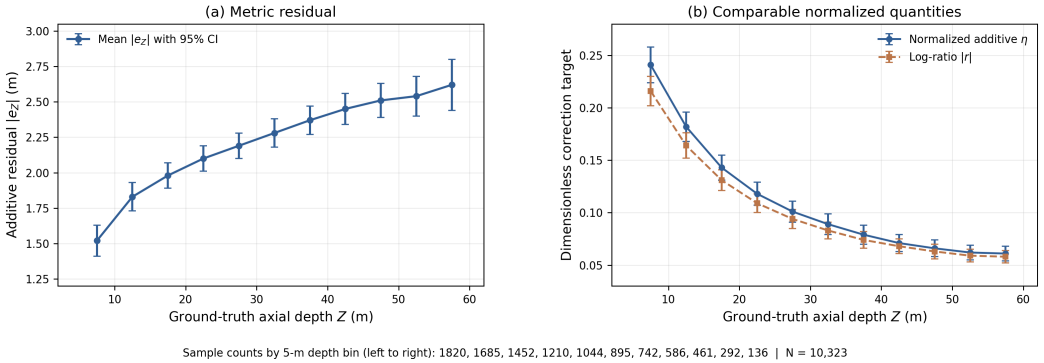


Figure 3. Geometry-correction targets on KITTI. (**a**) Mean $|e_Z|$ in meters versus ground-truth axial depth. (**b**) Dimensionless $\eta = |e_Z|/\hat{Z}_{\text{geo}}$ and $|r|$. Error bars are 95% confidence intervals. Counts under the panels sum to 10,323 objects across eleven 5 m bins.

Table 13. Central-depth MAE (m) on KITTI Car validation. Results describe individual checkpoints; lower is better. Bins use ground-truth axial depth: $(0, 20]$, $(20, 40]$, and $(40, \infty)$ m.

Method	0–20 m	20–40 m	>40 m	Overall
MonoDGP [15]	0.62	1.28	2.71	1.36
method(1): SPRR	0.61	1.23	2.56	1.31
method(2): SPRR + GSAU	0.60	1.20	2.48	1.28
method(3): RGC-Mono (ordinal OVCC)	**0.59**	**1.17**	**2.39**	**1.24**

Stabilization sensitivity. Table 14 varies ε_h at inference for the original ordinal checkpoint. Moderate AP3D and far-range MAE are unchanged at the displayed precision for $10^{-6} \le \varepsilon_h \le 10^{-3}$ pixels. At 10^{-1} pixels, AP3D decreases by 0.05 points and far-range MAE increases by 0.02 m. The last column evaluates the analytical perturbation $\varepsilon_h/(5 + \varepsilon_h)$ for a 5-pixel box.

Table 15 reports the Pearson correlation between the absolute correction target and ground-truth depth. The additive metric error exhibits a stronger distance dependence, whereas the log-ratio residual is substantially less correlated with depth.

Uncertainty and visibility analysis. GSAU models the reliability of the geometric anchor through uncertainty estimates tied to the underlying geometric variables.

Table 14. Sensitivity to the image-height stabilization constant for the original ordinal-OVCC checkpoint on KITTI validation (Car). The last column is the exact analytical relative depth perturbation at a projected height of 5 pixels; the other two columns report detection/depth evaluation results.

ε_h (Pixels)	Val Moderate $AP_{3D}\|_{R40}$ (%) ↑	>40 m Depth MAE (m) ↓	Perturbation at 5 px (%)
10^{-6}	23.31	2.39	0.00002
10^{-4}	23.31	2.39	0.00200
10^{-3} (default)	**23.31**	**2.39**	**0.02000**
10^{-2}	23.30	2.39	0.19960
10^{-1}	23.26	2.41	1.96078

Table 15. Pearson association between absolute correction magnitude and ground-truth axial depth on KITTI training. Smaller absolute correlation indicates weaker linear depth dependence.

Correction Target	Absolute Pearson Correlation with Depth	Interpretation
Additive geometry error e_Z	0.36	Metric residual retains depth-scale dependence
Log projection ratio r	**0.08**	Relative correction is substantially less coupled to distance

Definition of the baseline uncertainty and correlation protocol. The term *baseline uncertainty* refers specifically to the object-level central-depth standard deviation $\sigma_{d,i}^{\text{base}}$ output by the original MonoDGP depth-uncertainty head [15]. This head accompanies MonoDGP's heteroscedastic metric-depth regression; its uncertainty score is $u_i^{\text{base}} = \sigma_{d,i}^{\text{base}}$ (in meters). In contrast, GSAU uses the combined *log-depth* standard deviation from Equations (14), (18), and (19):

$$u_i^{\text{GSAU}} = \sigma_{d,i} = \sqrt{\sigma_{H,i}^2 + \sigma_{h,i}^2 + \sigma_{r,i}^2}.$$

The GSAU score is dimensionless because its variance components are defined in log depth; the baseline score is expressed in metric depth. For each matched Car detection i from model m on KITTI validation, we associate its uncertainty with the same model's absolute log-central-depth error, $e_i^{(m)} = |\log \hat{Z}_i^{(m)} - \log Z_i^{\text{gt}}|$, and compute

$$\rho_s^{(m)} = \text{Spearman}\left(\{u_i^{(m)}\}_{i=1}^{N_m}, \{e_i^{(m)}\}_{i=1}^{N_m}\right).$$

Both methods follow the same validation split, Car selection, and prediction-to-ground-truth matching protocol; each uncertainty score is paired with its corresponding model prediction. The resulting rank correlations are 0.43 for the original MonoDGP metric-depth uncertainty and 0.56 for GSAU from the original ordinal-OVCC RGC-Mono checkpoint. These coefficients describe depth-error ranking. Predictive-distribution calibration requires a separate assessment. They are also distinct from the Pearson error-dependence diagnostics reported below and the geometric-anchor disagreement statistic below.

As a complementary diagnostic, we define the relative depth disagreement between the calibrated depth and the geometric anchor as

$$d_{\text{rel}} = \frac{|\hat{Z} - \hat{Z}_{\text{geo}}|}{\hat{Z}_{\text{geo}} + \varepsilon_Z}.$$

Here, $\varepsilon_Z = 10^{-6}$ meters stabilizes a metric-depth denominator and is distinct from the dimensionless variance constant ε_v and pixel-valued height constant ε_h.

The statistic d_{rel} measures the model's correction magnitude. We compare it with ground-truth 3D-center error

$$E_{\text{3D}} = \|\hat{\mathbf{C}} - \mathbf{C}^{\text{gt}}\|_2,$$

using geometric box centers in the rectified KITTI frame. Figure 4 summarizes 11,315 matched detections by disagreement decile. The object-level Spearman correlations are 0.31 with 3D-center error and 0.25 with absolute log-depth error, indicating weak positive associations.

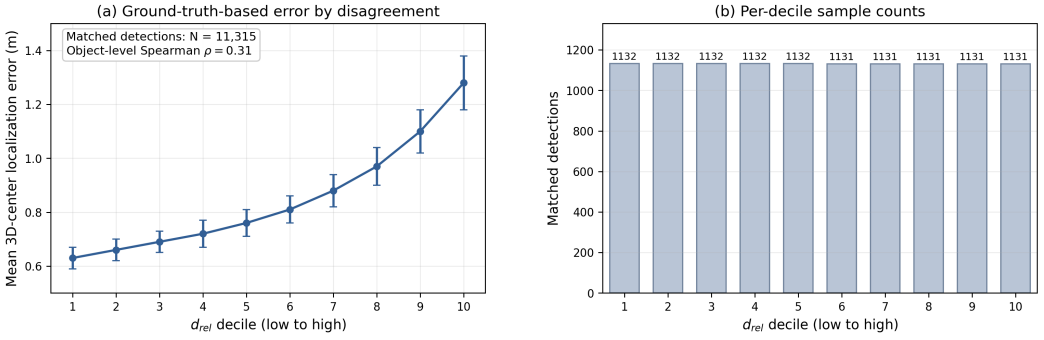


Figure 4. Geometric disagreement and localization error. (**a**) Mean 3D geometric-center error over ten d_{rel} deciles, with 95% confidence intervals. The object-level Spearman correlation is 0.31. (**b**) Counts per decile, totaling 11,315 matched detections. The panels summarize correction magnitude and ground-truth localization error.

For the depth-error breakdown below, only the three known KITTI occlusion states (0–2) are treated as ordered qualitative groups; label 3 (unknown) is not included in any severity bin.

Occlusion-stratified ablation of OVCC. To separate the contribution of orientation–visibility conditioning from the shared SPRR and GSAU components, Table 16 compares method(2) (SPRR + GSAU, without OVCC) against method(3) (SPRR + GSAU with the original ordinal OVCC). These two configurations have the same geometric anchor, log-ratio depth correction, GSAU loss, pretrained MonoDGP initialization, 250-epoch additional training budget, and KITTI validation evaluation protocol; the change is the introduction of the OVCC calibration pathway. We use the KITTI ground-truth occlusion labels 0 (fully visible), 1 (partly occluded), and 2 (largely occluded); label 3 denotes unknown visibility and is excluded from ordered severity groups. For each group o, the reported metric is the camera-frame axial-depth mean absolute error

$$\mathrm{MAE}_Z^{(o)} = \frac{1}{N_o} \sum_{i \in \mathcal{S}_o} \left| \hat{Z}_i - Z_i^{\mathrm{gt}} \right|, \qquad o \in \{0, 1, 2\},$$

where $\mathcal{S}_o$ contains the matched Car detections assigned to the group under the same evaluation and matching rules. The subgroup MAEs below are single-checkpoint results. MonoDGP is included to retain the context of the earlier full-model comparison; the method(2)-to-method(3) contrast is the relevant ablation for OVCC.

Table 16. OVCC contribution by occlusion state. KITTI Car validation axial-depth MAE (m), single checkpoints. Method(2) uses SPRR + GSAU; method(3) adds ordinal OVCC. Unknown state 3 is excluded. Reductions compare method(2) with method(3).

Method	Fully Visible (0) ↓	Partly Occluded (1) ↓	Largely Occluded (2) ↓
MonoDGP (prior comparison)	1.10	1.73	2.61
method(2): SPRR + GSAU (w/o OVCC)	1.07	1.66	2.49
method(3): SPRR + GSAU + OVCC (ordinal)	**1.05**	**1.58**	**2.33**
MAE reduction from OVCC (m)	0.02	0.08	0.16
Relative reduction from OVCC (%)	1.87	4.82	6.43

Adding ordinal OVCC reduces MAE by 0.02 m (1.87%) for visible objects, 0.08 m (4.82%) for partial occlusion, and 0.16 m (6.43%) for large occlusion. The full-model differences against MonoDGP are 0.05, 0.15, and 0.28 m. These subgroup measurements use the ordinal checkpoint; Figure 2 illustrates both improvements and failures under difficult observations.

Empirical assessment of the diagonal-covariance approximation. Table 17 summarizes the empirical error dependencies and omitted covariance contributions. We evaluate the neglected error dependencies in Equations (14) and (19) using the original ordinal-OVCC checkpoint and matched Car detections on KITTI validation. Ground-truth camera-frame axial depth Z^{gt} defines three distance intervals: 0–20 m, 20–40 m, and beyond 40 m. For each matched object, we compute the log-height errors e_H and e_h, the geometric-anchor error $e_g = e_H - e_h$, and the ratio-correction error $e_r = \mu_r - r^*$ using the definitions in Section 3.3. We report Pearson correlations $\rho_{Hh} = \mathrm{Corr}(e_H, e_h)$ and $\rho_{gr} = \mathrm{Corr}(e_g, e_r)$. The relative magnitude of the covariance omitted in Equation (14) is

$$D_g = \frac{2|\mathrm{Cov}(e_H, e_h)|}{\mathrm{Var}(e_H) + \mathrm{Var}(e_h)} \times 100\%.$$

This quantity normalizes omitted covariance by the empirical diagonal variance sum and characterizes second-order dependence within each depth range.

Overall, $\rho_{Hh} = 0.08$ and $D_g = 7.4\%$. Beyond 40 m, these increase to 0.13 and 12.0%, indicating a larger omitted height-error covariance at long range. The anchor–correction correlations are negative (-0.11 overall and -0.17 beyond 40 m), consistent with partial compensation of anchor error by the ratio branch.

Table 17. Empirical error dependence on KITTI Car validation with ordinal OVCC. ρ_{Hh} and ρ_{gr} are Pearson correlations. D_g is the normalized omitted height covariance, and $|\rho_{gr}| \times 100\%$ bounds the normalized anchor–correction covariance contribution.

Ground-Truth Depth	ρ_{Hh}	D_g (%)	ρ_{gr}	D_{gr} Bound (%)
0–20 m	0.04	3.7	−0.06	≤ 6
20–40 m	0.07	6.5	−0.10	≤ 10
>40 m	0.13	12.0	−0.17	≤ 17
Overall	0.08	7.4	−0.11	≤ 11

Anchor–residual covariance and the independence diagnostic. With $s_g^2 = \mathrm{Var}(e_g)$ and $s_r^2 = \mathrm{Var}(e_r)$, the omitted empirical covariance contribution relative to the diagonal error-variance sum satisfies

$$D_{gr} = \frac{2|\mathrm{Cov}(e_g, e_r)|}{s_g^2 + s_r^2} = |\rho_{gr}| \frac{2 s_g s_r}{s_g^2 + s_r^2} \leq |\rho_{gr}|.$$

The last column of Table 17 follows directly from the measured correlations: it bounds D_{gr} by 11% overall and 17% beyond 40 m. Since the observed correlations are negative, simply summing the corresponding *empirical* component-error variances would omit a negative correction and could overstate their total. These are bounds on empirical component-error variance. The predicted variance heads are jointly trained, so distributional calibration and conditional dependence require separate evaluation.

5. Limitations and Future Directions

Geometric-anchor dependence. The anchor depends on estimated object height, projected box height, and calibration. With $\tau = 1$, the bounded correction factor lies in (e^{-1}, e). Projection errors outside that range exceed the correction head's representational capacity. Adaptive bounds and alternative anchors may extend the range of recoverable errors.

Incomplete observations. Occlusion, truncation, and small distant projections reduce the geometric evidence available from one image. Ordinal RGC-Mono's depth MAE remains 2.39 m beyond 40 m, versus 2.71 m for MonoDGP. The long-range and occlusion failures in Figure 2e,f increase depth error relative to the baseline, showing that learned calibration can also degrade individual predictions.

Uncertainty approximation and attribution. The omitted height-error covariance contribution is 7.4% overall and 12.0% beyond 40 m. Anchor–correction dependence also affects variance addition, and the composite likelihood identifies the total variance more directly than its allocation. The GSAU ablation jointly introduces variance heads, auxiliary losses, and regression weighting; separating these contributions requires additional controls. Covariance-aware propagation and predictive-distribution calibration are useful extensions.

Evaluation coverage. The experiments cover KITTI Car. Categorical OVCC is evaluated on validation, while the test result belongs to ordinal OVCC. Pedestrian and Cyclist evaluation, camera shifts, and cross-dataset transfer require dedicated training and matched controls because object dimensions, articulation, and visibility statistics differ.

Representation and training variability. The capacity-matched SPRR–additive mean difference is 0.10 points with a confidence interval spanning zero. The algebraic scale-normalization property therefore has a more direct basis than an isolated detection advantage. Three training trials give positive full-model differences against matched MonoDGP,

with exact sign-flip $p = 0.25$. More trials and per-image AP resampling would improve estimates of training and evaluation variability.

6. Conclusions

RGC-Mono combines a perspective-derived anchor, SPRR log-ratio correction, GSAU training-time variance weighting, and OVCC orientation–visibility conditioning within MonoDGP. The log ratio expresses multiplicative projection discrepancy in a dimensionless variable, while the geometric anchor preserves the connection between image size and metric depth.

With a shared pretrained checkpoint and 250 additional epochs, three KITTI Car validation runs yield Moderate $AP_{3D|R40}$ of $22.91 \pm 0.14\%$ for matched MonoDGP, $23.26 \pm 0.12\%$ for ordinal RGC-Mono, and $23.36 \pm 0.08\%$ for categorical RGC-Mono. The integrated configurations improve the mean by 0.35 and 0.45 percentage points. The original ordinal checkpoint obtains 19.49% on the KITTI test under its reported protocol.

The ablations characterize the result: GSAU raises mean Moderate AP3D by 0.29 points over SPRR, and ordinal OVCC reduces depth MAE by 0.02/0.08/0.16 m across visible/partly/largely occluded groups. The capacity-matched SPRR–additive interval spans zero, placing the detection evidence primarily with the integrated configuration. Categorical OVCC preserves KITTI's unknown-state semantics and offers a compact extension of the calibrated geometry pathway.

Author Contributions: Conceptualization, Y.Z.; methodology, A.X.; software, Y.Z.; validation, A.X.; formal analysis, Y.Z.; investigation, A.X.; resources, Y.Z.; data curation, Y.Z.; writing—original draft preparation, Y.Z.; writing—review and editing, Y.Z.; visualization, Y.Z.; supervision, A.X.; project administration, Y.Z. 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 analyzed in this study is available from the KITTI Vision Benchmark Suite at https://www.cvlibs.net/datasets/kitti/, accessed on 9 November 2026. The study-specific predictions, analysis outputs, code, and model checkpoints are not shared.

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

References

1. 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), Virtual Event, 11–17 October 2021.
2. Liu, X.; Xue, N.; Wu, T. Learning Auxiliary Monocular Contexts Helps Monocular 3D Object Detection. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Virtual Event, 22 February–1 March 2022.
3. Huang, K.-C.; Wu, T.-H.; Su, H.-T.; Hsu, W.H. MonoDTR: Monocular 3D Object Detection with Depth-Aware Transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022.
4. Li, Z.; Qu, Z.; Zhou, Y.; Liu, J.; Wang, H.; Jiang, L. Diversity Matters: Fully Exploiting Depth Clues for Reliable Monocular 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022.
5. Peng, L.; Wu, X.; Yang, Z.; Liu, H.; Cai, D. DID-M3D: Decoupling Instance Depth for Monocular 3D Object Detection. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022.
6. Kumar, A.; Brazil, G.; Corona, E.; Parchami, A.; Liu, X. DEVIANT: Depth EquiVarIAnt NeTwork for Monocular 3D Object Detection. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022.
7. Qin, Z.; Li, X. MonoGround: Detecting Monocular 3D Objects from the Ground. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022.
8. Zhang, R.; Qiu, H.; Wang, T.; Guo, Z.; Cui, Z.; Qiao, Y.; Li, H.; Gao, P. MonoDETR: Depth-Guided Transformer for Monocular 3D Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023.

9. Jia, J.; Li, Z.; Shi, Y. MonoUNI: A Unified Vehicle and Infrastructure-Side Monocular 3D Object Detection Network with Sufficient Depth Clues. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, 10–16 December 2023; Volume 36.
10. Wu, Z.; Gan, Y.; Wang, L.; Chen, G.; Pu, J. MonoPGC: Monocular 3D Object Detection with Pixel Geometry Contexts. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), London, UK, 29 May–2 June 2023.
11. Liu, X.; Zheng, C.; Cheng, K.B.; Xue, N.; Qi, G.-J.; Wu, T. Monocular 3D Object Detection with Bounding Box Denoising in 3D by Perceiver. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023.
12. Yan, L.; Yan, P.; Xiong, S.; Xiang, X.; Tan, Y. MonoCD: Monocular 3D Object Detection with Complementary Depths. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024.
13. Peng, L.; Xu, J.; Cheng, H.; Yang, Z.; Wu, X.; Qian, W.; Wang, W.; Wu, B.; Cai, D. Learning Occupancy for Monocular 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024.
14. Wu, Z.; Gan, Y.; Wu, Y.; Wang, R.; Wang, X.; Pu, J. FD3D: Exploiting Foreground Depth Map for Feature-Supervised Monocular 3D Object Detection. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Vancouver, BC, Canada, 20–27 February 2024; Volume 38.
15. Pu, F.; Wang, Y.; Deng, J.; Yang, W. MonoDGP: Monocular 3D Object Detection with Decoupled-Query and Geometry-Error Priors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025.
16. Kumar, A.; Guo, Y.; Zhang, Z.; Huang, X.; Ren, L.; Liu, X. CHARM3R: Towards Unseen Camera Height Robust Monocular 3D Detector. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–23 October 2025.
17. Hu, Z.; Li, D.; Ma, X.; Tang, S.; Li, X.; Yang, W.; Duan, L.-Y. Adaptive Dual Uncertainty Optimization: Boosting Monocular 3D Object Detection under Test-Time Shifts. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–23 October 2025.
18. Wang, Y.; Zhao, Y.; Pu, F.; Yang, X.; Tang, Y.; Chen, X.; Yang, W. SPAN: Spatial-Projection Alignment for Monocular 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Denver, CO, USA, 3–7 June 2026; pp. 40762–40771.
19. Zhang, Z.; Kumar, A.; Liu, X. Towards Intrinsic-Aware Monocular 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Denver, CO, USA, 3–7 June 2026; pp. 40739–40750.
20. Park, D.; Ambrus, R.; Guizilini, V.; Li, J.; Gaidon, A. Is Pseudo-Lidar Needed for Monocular 3D Object Detection? In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Virtual Event, 11–17 October 2021.
21. 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), Virtual Event, 19–25 June 2021.
22. Peng, L.; Liu, F.; Yu, Z.; Yan, S.; Deng, D.; Yang, Z.; Liu, H.; Cai, D. Lidar Point Cloud Guided Monocular 3D Object Detection. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022.
23. Lian, Q.; Xu, Y.; Yao, W.; Chen, Y.; Zhang, T. Semi-Supervised Monocular 3D Object Detection by Multi-View Consistency. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022.
24. Li, Z.; Chen, Z.; Li, A.; Fang, L.; Jiang, Q.; Liu, X.; Jiang, J. Unsupervised Domain Adaptation for Monocular 3D Object Detection via Self-Training. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022.
25. Su, Y.; Di, Y.; Zhai, G.; Manhardt, F.; Rambach, J.; Busam, B.; Stricker, D.; Tombari, F. OPA-3D: Occlusion-Aware Pixel-Wise Aggregation for Monocular 3D Object Detection. *IEEE Robot. Autom. Lett.* **2023**, *8*, 1327–1334. [CrossRef]
26. Lin, H.; Zhang, Y.; Niu, S.; Cui, S.; Li, Z. MonoTTA: Fully Test-Time Adaptation for Monocular 3D Object Detection. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024.
27. Skvrna, J.; Neumann, L. MonoSOWA: Scalable Monocular 3D Object Detector Without Human Annotations. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–23 October 2025.
28. Jung, J.; Kim, S.; Kim, J.U. MonoSAOD: Monocular 3D Object Detection with Sparsely Annotated Label. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Denver, CO, USA, 3–7 June 2026.
29. Yang, Y.-H.; Piccinelli, L.; Segu, M.; Li, S.; Huang, R.; Fu, Y.; Pollefeys, M.; Blum, H.; Bauer, Z. 3D-MOOD: Lifting 2D to 3D for Monocular Open-Set Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–23 October 2025.
30. Zhou, Y.; Zhu, H.; Liu, Q.; Chang, S.; Guo, M. MonoATT: Online Monocular 3D Object Detection with Adaptive Token Transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023.

31. Xu, J.; Peng, L.; Cheng, H.; Li, H.; Qian, W.; Li, K.; Wang, W.; Cai, D. MonoNeRD: NeRF-like Representations for Monocular 3D Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023.
32. Li, Z.; Xu, X.; Lim, S.; Zhao, H. UniMODE: Unified Monocular 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024.
33. Ranasinghe, Y.; Hegde, D.; Patel, V.M. MonoDiff: Monocular 3D Object Detection and Pose Estimation with Diffusion Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024.
34. Park, H.; Xiong, P.; Ruan, X.; Jia, D.; Yu, P.; Tang, W. RARE: Learn to RAnk and REtrieve for Monocular 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Denver, CO, USA, 3–7 June 2026.
35. Zhang, Z.; Kumar, A.; Ganesan, G.C.; Liu, X. Unleashing the Power of Chain-of-Prediction for Monocular 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Denver, CO, USA, 3–7 June 2026.
36. 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), Virtual Event, 14–19 June 2020.
37. Brazil, G.; Pons-Moll, G.; Liu, X.; Schiele, B. Kinematic 3D Object Detection in Monocular Video. In Proceedings of the European Conference on Computer Vision (ECCV), Virtual Event, 23–28 August 2020.
38. 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), Virtual Event, 19–25 June 2021.
39. Luo, S.; Dai, H.; Shao, L.; Ding, Y. M3DSSD: Monocular 3D Single Stage Object Detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual Event, 19–25 June 2021.
40. Ma, X.; Zhang, Y.; Xu, D.; Zhou, D.; Yi, S.; Li, H.; Ouyang, W. Delving into Localization Errors for Monocular 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual Event, 19–25 June 2021.
41. Shi, X.; Ye, Q.; Chen, X.; Chen, C.; Chen, Z.; Kim, T.-K. Geometry-Based Distance Decomposition for Monocular 3D Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Virtual Event, 11–17 October 2021.
42. Li, Y.; Chen, Y.; He, J.; Zhang, Z. Densely Constrained Depth Estimator for Monocular 3D Object Detection. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022.
43. Wang, T.; Zhu, X.; Pang, J.; Lin, D. Probabilistic and Geometric Depth: Detecting Objects in Perspective. In Proceedings of the 5th Conference on Robot Learning (CoRL), London, UK, 8–11 November 2021; Proceedings of Machine Learning Research, 2022; Volume 164, pp. 1475–1485.
44. Wang, T.; Pang, J.; Lin, D. Monocular 3D Object Detection with Depth from Motion. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022.
45. Zhu, M.; Ge, L.; Wang, P.; Peng, H. MonoEdge: Monocular 3D Object Detection Using Local Perspectives. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2–7 January 2023.
46. Eigen, D.; Puhrsch, C.; Fergus, R. Depth Map Prediction from a Single Image Using a Multi-Scale Deep Network. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 8–13 December 2014; Volume 27, pp. 2366–2374.
47. Fu, H.; Gong, M.; Wang, C.; Batmanghelich, K.; Tao, D. Deep Ordinal Regression Network for Monocular Depth Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 2002–2011.
48. Bhat, S.F.; Birkl, R.; Wofk, D.; Wonka, P.; Müller, M. ZoeDepth: Zero-Shot Transfer by Combining Relative and Metric Depth. *arXiv* **2023**, arXiv:2302.12288.
49. Zhu, R.; Wang, C.; Song, Z.; Liu, L.; He, J.; Deng, J.; Zhang, T.; Zhang, Y. ScaleDepth: Decomposing Metric Depth Estimation into Semantic-Aware Scale Prediction and Adaptive Relative Depth Estimation. *IEEE Trans. Circuits Syst. Video Technol.* **2026**, *36*, 9547–9560. [CrossRef]
50. Kendall, A.; Gal, Y. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
51. Kendall, A.; Gal, Y.; Cipolla, R. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7482–7491.
52. Lu, Y.; Ma, X.; Yang, L.; Zhang, T.; Liu, Y.; Chu, Q.; He, T.; Li, Y.; Ouyang, W. GUPNet++: Geometry Uncertainty Propagation Network for Monocular 3D Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 900–915. [CrossRef] [PubMed]
53. 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.

54. Jiang, X.; Jin, S.; Zhang, X.; Shao, L.; Lu, S. MonoMAE: Enhancing Monocular 3D Detection through Depth-Aware Masked Autoencoders. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 10–15 December 2024; Volume 37.
55. Park, H.; Xiong, P.; Yu, P.; Tang, W. Modeling and Learning Multiple Hypotheses for Monocular 3D Object Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Tucson, AZ, USA, 6–10 March 2026; pp. 7040–7050.

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.