

Research On Target Visual Depth Estimation In Highly Reflective Bin-Picking Scenarios

Yunhui Zhang, Qi Yang*, Minghao Li, Hui Han*, and Yupu Zhang

* School of Mechanical Engineering, Shenyang Ligong University, Shenyang 110159, China

* Corresponding author. E-mail: yangqi@sylu.edu.cn; 34352350@qq.com

Received: August 18, 2026; Accepted: September 20, 2026

Accurate monocular metric depth perception is crucial for complex robotic tasks such as bin-picking. However, in real industrial scenes, specular reflections from metal surfaces and mutual occlusions cause severe feature distortions and scale inconsistencies. To address this issue, we propose a Geometric-Radiometric Collaborative Framework guided by specular failure priors for robust absolute metric reconstruction under strong specular reflections. We decouple canonical depth reconstruction from global metric calibration and use affine parameters s and t to map scale-independent canonical depth to metric depth. In addition, a specular-aware energy constraint combining mask guidance and image gradients suppresses reflection-induced depth collapse and local distortions. Experiments on the XYZ-IBD dataset demonstrate that our method consistently outperforms the evaluated baseline models. In highly reflective metal regions, the framework lowers the root mean square error (RMSE) to 97.0 mm, increases the δ_1 accuracy to 91.39%, and reduces the Local Surface Smoothness (LSS) metric to 3.04. These results provide a reliable basis for 3D reconstruction in vision-based robotic bin-picking.

Keywords: Monocular depth estimation; Absolute scale recovery; Energy functional optimization; Specular reflection;

Robotic bin-picking

© The Author(s). This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY 4.0\)](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are cited.

http://dx.doi.org/10.6180/jase.202701_36.017

1. Introduction

Accurate 6-Degree-of-Freedom (6DoF) object pose estimation fundamentally relies on the metric accuracy of the underlying 3D geometric perception [1, 2]. When facing metal parts commonly found in industrial manufacturing, traditional active sensing technologies (e.g., structured light) suffer from substantial missing-depth regions due to strong specular radiance degradation (with an invalid pixel rate as high as 46.0% on the XYZ-IBD dataset [3]). Conventional scale-agnostic depth methods, such as MiDaS [4], Omnidata [5], and LeReS [6], cannot directly recover absolute metric scale. In contrast, vision-only Monocular Metric Depth Estimation (MMDE) provides physical depth measurements required for robotic operations and therefore offers a potential way to overcome this limitation.

However, both early depth classification methods like AdaBins and existing state-of-the-art MMDE methods (e.g., UniDepthV2 [7]) reveal dual flaws when transferred to high-fidelity industrial perception [8, 9]: On the one hand, large vision models tend to develop an “implicit focal-length prior” during pre-training in closed scenes. During deployment, explicitly introducing camera intrinsics during inference may conflict with the implicit focal-length prior learned during pre-training, resulting in a systematic “Double Scaling Conflict” and causing global scale drift and local depth distortion in physical space. Another limitation arises from the Lambertian surface assumption adopted by many existing algorithms [10, 11]. Real industrial bin-picking scenes often contain highly reflective metal parts with dynamic specular highlights and severe mutual occlusions. These effects violate the photometric consistency

assumption. As a result, depth predictions may contain large invalid regions, local extremum distortions, and fragmented geometric structures. Such perceptual degradation further reduces the reliability of downstream vision-only registration and grasp pose estimation.

To address the challenges in highly reflective bin-picking, we propose a Geometric-Radiometric Collaborative Framework to reshape monocular metric perception. The main contributions are:

- **Implicit Scale Conflict Analysis:** We systematically benchmark radiance degradation in existing MMDE methods and analyze a potential source of physical scale collapse induced by an “implicit focal-length prior”.
- **Intrinsic-Free Scale Decoupling:** We decouple Canonical Depth generation from global calibration. By relying solely on a known physical distance from a single frame (e.g., a bin edge) for rapid median scale alignment, we reduce the Double Scaling Conflict without requiring explicit camera intrinsics.
- **Metal-Aware Collaborative Energy Constraint:** Innovatively transforming active sensor Specular Mask into meta-supervisory signals, we design an anisotropic energy functional. This forces the network to rely on occlusion contours for non-local smoothness, effectively suppressing specular extremum distortions without manual annotation.
- **Industrial-Grade Validation:** On the XYZ-IBD dataset, our framework achieves a state-of-the-art RMSE of 97.0 mm in core metal regions, increases δ_1 accuracy to 91.39%, and significantly reduces the Local Surface Smoothness (LSS) error to 3.04, advancing vision-only robotic grasping.

2. Methods

2.1. Problem Definition and Overall Algorithm Architecture

Deploying expensive 3D sensors in automated bin-picking is often hindered by strict payload and cost limits. While vision-only absolute metric depth estimation offers a lightweight alternative, the direct application of ViT-based foundation models to industrial scenarios is hindered by a severe domain gap. Specifically, the lack of conventional semantic priors (e.g., room structures) causes global absolute scale drift, while strong specular reflections on metal parts disrupt local photometric and geometric continuity, inducing severe local topological distortions.

From the perspective of metric reconstruction, the objective is to infer a dense depth field with physically meaningful scale from a single RGB image. Let the input image be I and the desired metric depth map be D . In industrial bin-picking scenes, prediction errors mainly arise from two coupled factors. The first is global metric inconsistency, where the reconstructed scene is incorrectly scaled or shifted in physical space. The second is local structural inconsistency, where specular highlights, textureless surfaces, and mutual occlusions introduce abnormal depth gradients, holes, and non-physical surface fluctuations. Jointly optimizing these two error sources within a single regression objective may entangle global scale adaptation with local topology reconstruction, leading to unstable optimization under strong radiometric disturbances. This coupling is particularly undesirable in industrial scenes, where global metric calibration and local reflection-induced distortions originate from different physical factors and should therefore be handled separately.

Based on this observation, we formulate metric depth recovery as a decoupled reconstruction problem. The network first reconstructs a scale-independent Canonical Depth representation that preserves relative geometric structure, while the global metric relationship is recovered through a lightweight affine transformation parameterized by a scale factor s and an offset t . This design allows the high-dimensional visual representation to remain within the geometric manifold learned during large-scale pre-training, while the low-dimensional affine parameters absorb scene-dependent metric variations. As a result, the network can preserve stable geometric inference while adapting the global metric scale with minimal disturbance to the pre-trained feature space.

Local geometric degradation is handled separately through radiometry-aware structural refinement. Since highly reflective pixels provide unreliable appearance cues, we extract a specular failure prior to identify regions where conventional photometric evidence becomes unreliable. Together with image-gradient information, this prior guides an energy-based refinement process that strengthens geometric fidelity, suppresses reflection-induced depth discontinuities, and preserves genuine object boundaries. The resulting framework follows a coarse-to-fine principle: global affine calibration establishes the metric manifold, while local collaborative energy constraints repair reflection-sensitive topology, forming the basis of the overall architecture illustrated in Fig. 1.

To resolve these bottlenecks, we propose a two-stage fine-tuning architecture coupling global physical calibration with local geometric repair. Built upon a frozen ViT-

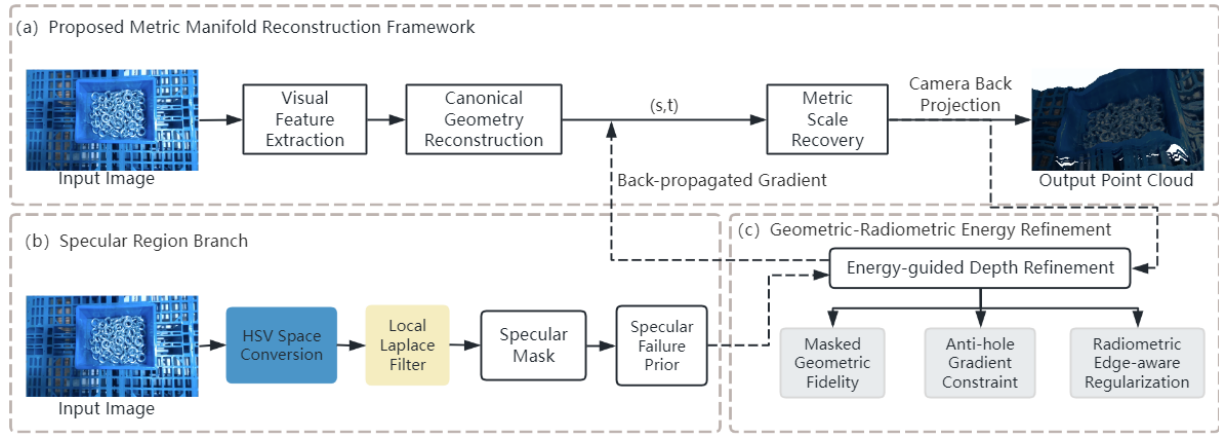


Fig. 1. The proposed Geometric-Radiometric Collaborative Framework.

Large encoder to retain robust generalized features, our framework achieves high-precision metric depth prediction through three core components: an adaptive specular mask extraction mechanism, an affine-based global scale alignment strategy, and an Anti-hole Gradient Constraint.

2.2. Energy Functional Methods for Depth Field Optimization

To eliminate topological breakage caused by highlight interference, traditional local filters or Markov Random Fields (MRFs) often blur real physical object edges while smoothing out distortions when dealing with depth discontinuities [12, 13]. Therefore, we discard the conventional global smoothness assumption and propose an intrinsic-free scale-decoupling mechanism together with a geometric-radiometric collaborative energy functional adaptively guided by image gradients. Given an initial depth prediction D_{init} and an input image I , the collaborative energy functional E proposed in this paper is defined on a continuous 2D image domain Ω as:

$$E(D) = \int_{\Omega} [E_{data}(D, D_{init}) + \lambda E_{collab}(\nabla D, \nabla I, M)] dx \quad (1)$$

Here, the “Specular Mask” M , derived from the physical limitations of active sensors, is specifically introduced as a prior. By minimizing this energy functional, the optimization downweights unreliable appearance cues in specular regions and places greater emphasis on high-confidence geometric boundaries for non-local smoothness propagation, thereby achieving high-fidelity reshaping of the absolute metric topology in difficult areas within a unified mathematical framework.

2.3. Robust Specular Mask Extraction Under Strong Illumination Interference

To mitigate local feature collapse caused by specular reflections, we propose an unsupervised specular mask extraction algorithm. Since illumination and color are highly coupled in RGB, we first convert the image to the HSV color space. According to the dichromatic reflection model [14], intense specular highlights cause severe local overexposure on the camera’s CMOS sensor. Consequently, these regions typically exhibit high Value (V) and low Saturation (S) values, regardless of the metal’s intrinsic color or global lighting fluctuations. Thus, the initial mask is defined as:

$$M_0(x, y) = \begin{cases} 1, & V(x, y) > \tau_v \text{ and } S(x, y) < \tau_s, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

Through empirical sweeps on the training set, we apply fixed domain-specific thresholds ($\tau_v = 170$, $\tau_s = 60$). Using fixed thresholds reduces the baseline drift that may occur with dynamic thresholding. Furthermore, to prevent misclassifying inherently bright silver metals as specular highlights, we complement the absolute thresholds with a local Laplacian high-frequency contrast constraint:

$$V_{local}(x, y) = V(x, y) - \mathbf{M}_{k \times k}(V)(x, y) \quad (3)$$

where $\mathbf{M}_{k \times k}$ represents a mean filtering operation. A pixel is classified as a specular anomaly only if $V_{local} > \tau_{local}$ (with $\tau_{local} = 15$). Finally, a 1–2-pixel morphological dilation is applied to cover surrounding halo distortions. This dual design combining absolute physical saturation priors with local relative contrast ensures highly reliable specular isolation with minimal computational overhead.

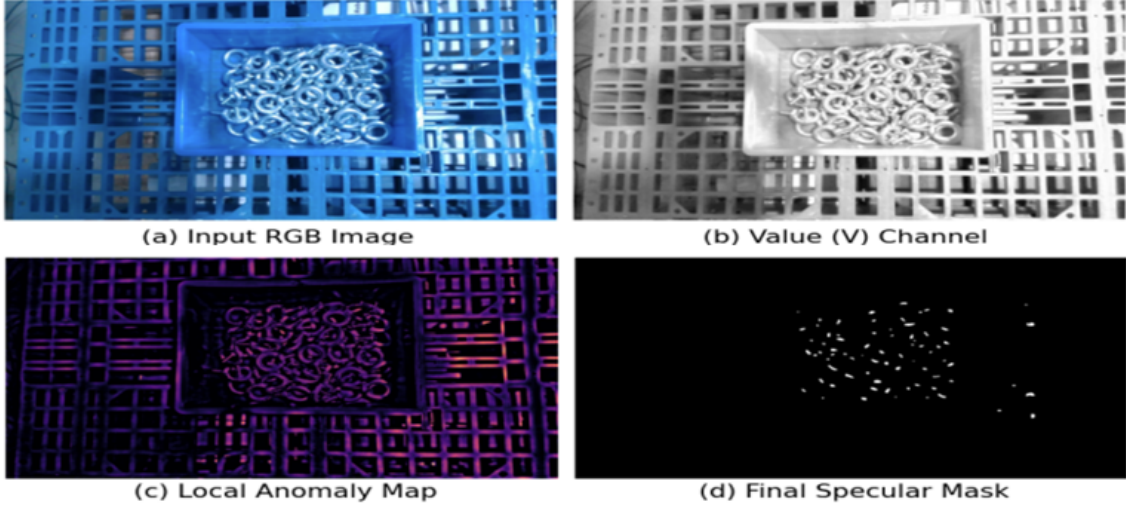


Fig. 2. Specular mask extraction process diagram. (a) Original RGB image with specular reflection; (b) Value (V) channel in the HSV color space; (c) Local anomaly map calculated through contrast constraints, highlighting regions with photometric surges; (d) The final robust specular mask after threshold processing and morphological dilation, which effectively isolates shiny metallic regions while excluding background noise.

2.4. Global Affine Alignment Mechanism Decoupling Camera Intrinsic

2.4.1. Mathematical Explanation of the Implicit Focal-Length Prior and Double Scaling Conflict

During pre-training on massive internet datasets without real camera intrinsics, vision foundation models learn perspective relationships and form a canonical implicit focal-length prior, denoted as f_{imp} . According to the standard pinhole camera model, the perspective projection relationship mapping 3D points (X, Y, Z) to 2D image coordinates (u, v) is:

$$Z \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \quad (4)$$

Without intrinsic input during zero-shot inference, the network output D_{net} represents a relative depth quantity implicitly conditioned on f_{imp} :

$$D_{\text{net}} = \phi(I | f_{\text{imp}}) \propto f_{\text{imp}} \cdot \frac{X_{\text{norm}}}{u - c_x} \quad (5)$$

where ϕ is the feature mapping function and X_{norm} is the object's physical size prior in canonical space. Existing explicit metric depth methods often attempt to recover the absolute scale by directly injecting the real focal length f_{true} into the network feature layer or forcing a post-scale. However, since the network's underlying Canonical Depth ϕ is already solidified by f_{imp} , explicitly introducing f_{true} during decoding causes a superposition effect of implicit priors and explicit parameters [15]:

$$D_{\text{out}} = \gamma \cdot f_{\text{true}} \cdot \phi(I | f_{\text{imp}}) \quad (6)$$

This forced coupling induces a non-linear scale drift dependent on the ratio $f_{\text{true}}/f_{\text{imp}}$, which we term the Double Scaling Conflict. In fixed-viewpoint, low-texture industrial scenarios, this intrinsic contradiction directly causes feature mapping failures and topological collapse of the depth field.

2.4.2. Equivalence Proof of Manifold Decoupling and Global Affine Alignment

To circumvent the distortion caused by the Double Scaling Conflict, we abandon explicit intrinsic injection and instead propose a global affine mechanism to decouple Canonical Depth prediction from physical scale mapping. Assuming the true physical depth is d_{GT} (determined by f_{true}) and the Canonical Depth is $d_{\text{pred}} = \phi(I | f_{\text{imp}})$, we apply learnable scale (s) and shift (t) factors outside the network:

$$d_{\text{aligned}}(x, y) = s \cdot d_{\text{pred}}(x, y) + t \quad (7)$$

To align d_{aligned} with d_{GT} , s acts as an implicit focal length compensator. Upon convergence, it is mathematically equivalent to:

$$s \approx k \cdot \frac{f_{\text{true}}}{f_{\text{imp}}} \quad (8)$$

where k absorbs absolute scene scale differences. Similarly, t compensates for systematic translations caused by principal point deviations:

$$t \approx \Delta Z (c_{x,\text{true}} - c_{x,\text{imp}}, c_{y,\text{true}} - c_{y,\text{imp}}) \quad (9)$$

The core advantage of this mechanism is that the linear affine mapping occurs after high-dimensional feature de-

coding. This allows the network ϕ to perform stable geometric inference within its native f_{imp} manifold, avoiding forced adaptation to f_{true} at the feature level. During deployment, s and t can be efficiently solved via the median ratio on a single reference frame, safely absorbing the scale conflict and thereby improving the stability of 3D reconstruction in reflective scenes.

2.5. Collaborative Energy Functional Optimization with Gradient Constraints

To mitigate depth collapse in specular regions, we formulate the loss function as a Geometric-Radiometric Collaborative Energy Functional:

$$E_{\text{total}} = \lambda_1 E_{\text{data}} + \lambda_2 E_{\text{grad}} + \lambda_3 E_{\text{smooth}} \quad (10)$$

Based on empirical grid search, we set $\lambda_1 = 1.0$, $\lambda_2 = 0.5$, and $\lambda_3 = 0.1$ to balance depth accuracy, local geometric continuity, and edge preservation.

We first define a spatial weight map $W(x, y)$ to emphasize pixels in specular metal regions:

$$W(x, y) = 1.0 + M_{\text{metal}}(x, y) \cdot (\omega_{\text{metal}} - 1.0) \quad (11)$$

where $M_{\text{metal}}(x, y) \in \{0, 1\}$ denotes the Specular Mask and ω_{metal} controls the additional weight assigned to these pixels. We set $\omega_{\text{metal}} = 3.0$ in our experiments. A second binary mask, $M_{\text{valid}}(x, y)$, indicates pixels with valid ground-truth depth and excludes invalid sensor measurements. The Masked Geometric Fidelity Term is then written as:

$$E_{\text{data}} = \frac{\sum_{x,y} W(x, y) \cdot M_{\text{valid}}(x, y) \cdot |d_{\text{aligned}}(x, y) - d_{\text{GT}}(x, y)|}{\sum_{x,y} W(x, y) \cdot M_{\text{valid}}(x, y)} \quad (12)$$

Here, d_{aligned} is the predicted metric depth after scale-shift alignment, and d_{GT} is the ground-truth depth. This term increases the contribution of specular regions while ignoring pixels without valid depth supervision.

To preserve local depth discontinuities while suppressing reflection-induced holes, we match the first-order gradients of the predicted and ground-truth depth maps:

$$E_{\text{grad}} = \frac{1}{N} \sum_{x,y} M_{\text{grad}}(x, y) \left(|\nabla_x d_{\text{aligned}}(x, y) - \nabla_x d_{\text{GT}}(x, y)| + |\nabla_y d_{\text{aligned}}(x, y) - \nabla_y d_{\text{GT}}(x, y)| \right) \quad (13)$$

where $M_{\text{grad}}(x, y)$ is a validity mask that retains only pixel pairs with valid depth values, and N is the number of valid gradient samples. The operators ∇_x and ∇_y denote first-order forward differences along the horizontal and vertical directions, respectively. For example,

$$\nabla_x d(x, y) = |d(x, y) - d(x + 1, y)| \quad (14)$$

We use direct first-order differences rather than additional smoothing-based gradient operators so that sharp depth transitions are not unnecessarily blurred.

To penalize unwarranted depth changes on textureless metal surfaces [16], we use the luminance gradient I_Y , computed from the RGB image, to provide an edge-aware structural prior. While the HSV space was employed earlier to isolate photometric specularities, the YUV space explicitly decouples luminance, providing a structural edge prior that is more robust to color noise:

$$E_{\text{smooth}} = \frac{1}{N} \sum_{x,y} |\nabla d_{\text{aligned}}| \cdot e^{-\|\nabla I_Y\|} \quad (15)$$

2.6. Closed-Loop Calibration in Practical Deployment

During deployment, the affine parameters s and t are solved via a lightweight, single-shot calibration. By extracting pixels from a known reference plane (e.g., a bin edge localized via mounting priors or robot kinematics) and comparing their median relative depth against the known physical distance, the global metric scale is calibrated using a median-based scale ratio, while the learned affine offset is retained. In contrast to conventional metric-depth approaches that rely on an accurate intrinsic matrix K , the proposed weak-prior-guided mechanism decouples metric mapping from complex multi-view geometry, thereby improving tolerance to calibration uncertainty and simplifying practical deployment.

3. Results and discussion

3.1. Experimental Preparation

The experiments are based on the challenging test set of the large-scale industrial bin-picking dataset XYZ-IBD. Unlike conventional indoor datasets, XYZ-IBD specifically captures the physical characteristics of real-world industrial environments. It covers a variety of textureless, stacked, highly reflective metal parts. The dense stacking of these industrial components inherently creates severe mutual occlusions and intricate multipath light reflections. Furthermore, the metallic surface finishes exhibit substantial variation in specular reflectance under active lighting. This optical degradation leads to large missing-depth regions in the ground truth provided by high-precision industrial-grade active optical sensors (e.g., Photoneo). During testing and training, the input image resolution was uniformly resized to 1280×720 to strictly match typical industrial camera sensing scenes.

In terms of network training, we use a Parameter-Efficient Fine-Tuning (PEFT) strategy [17]. A core design choice in this architecture is freezing most parameters of the

ViT-Large encoder. Having been pre-trained on internet-scale vision datasets, the encoder has already constructed a highly robust and generalized Canonical Depth, encompassing rich geometric priors such as perspective logic and occlusion contour awareness. Freezing the encoder parameters helps mitigate the “catastrophic forgetting” phenomenon that may occur when large vision models are fine-tuned on relatively small-scale, domain-specific industrial datasets. This preserves the encoder’s zero-shot feature extraction capability while allowing the lightweight decoder to adapt the local geometric representation to the highly reflective domain.

The model introduces affine parameters (s, t) to adapt to the depth distribution of the target domain. AdamW is used as the optimizer, with different learning rates assigned to the decoder feature layers and affine parameters. Training employs a cosine annealing learning rate scheduler and gradient clipping with a threshold of 0.5. Strong photometric disturbances on specular metal surfaces may produce large gradient variations, increasing the risk of numerical instability and gradient explosion. Gradient clipping is therefore applied before each parameter update to limit excessive gradient magnitudes and improve training stability. Performance is measured using standard AbsRel and RMSE. Crucially, to quantify non-physical high-frequency distortions, such as spike artifacts in specular regions, we further introduce the Local Surface Smoothness (LSS) metric:

$$LSS = \frac{1}{N} \sum_{x,y} |\nabla^2 D_{\text{pred}}(x, y)| \quad (16)$$

where ∇^2 is the discrete Laplacian operator. A lower LSS indicates a smoother, more physically coherent 3D surface.

Together, these metrics evaluate the reconstructed depth field from complementary perspectives. AbsRel measures the relative magnitude of depth errors, while RMSE emphasizes large absolute deviations that are particularly critical for metric robotic manipulation. In contrast, LSS focuses on local geometric regularity and is therefore sensitive to non-physical spikes and surface fluctuations induced by specular reflections. Accordingly, the following experiments report both conventional metric-depth errors and region-specific structural accuracy to assess the proposed framework under highly reflective industrial conditions.

3.2. Experimental Results

3.2.1. Quantitative Analysis

Table 1 compares the metric depth estimation performance across the full scene and the isolated metal region. Among all evaluated methods, the proposed Geometric-Radiometric Collaborative Framework achieves the best

performance across all reported metrics. For the full scene, the proposed method achieves an AbsRel of 0.0860 and an RMSE of 87.6 mm. In the metal region, the proposed method further achieves an AbsRel of 0.0948, an RMSE of 97.0 mm, and a δ_1 accuracy of 91.39%. Compared with the strongest baseline, UniDepthV2, the metal-region AbsRel and RMSE are reduced by approximately 18.1% and 12.9%, respectively, while the δ_1 accuracy increases from 84.98% to 91.39%. These results demonstrate that the proposed framework improves both overall metric accuracy and depth estimation robustness in highly reflective regions. The consistent gains across the full scene and the metal region further indicate that the improvement is not limited to global scale correction, but also extends to local depth reconstruction under strong specular interference. This behavior is consistent with the combined effects of affine-based scale alignment and specular-aware geometric-radiometric energy refinement.

Notably, the improvement is more pronounced in the metal region, indicating that the proposed framework effectively addresses reflection-induced degradation rather than merely reducing global prediction errors. The affine-based scale alignment improves global metric consistency, while the specular-aware geometric-radiometric energy constraints suppress local depth distortions in highly reflective regions.

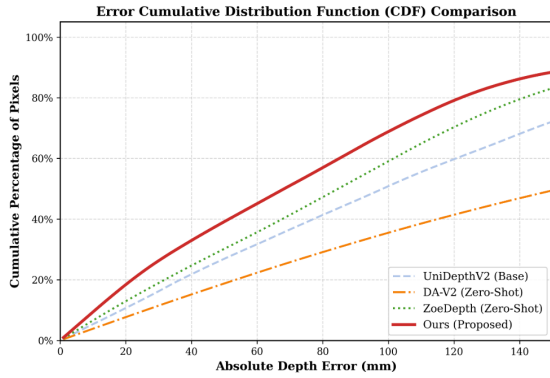
Error Cumulative Distribution Analysis: Figure 3 shows the cumulative distribution of pixel-wise absolute depth errors in the core metal regions. At an absolute error threshold of 150 mm, approximately 72%, 49%, and 83% of pixels for UniDepthV2, DA-V2, and ZoeDepth, respectively, fall within this tolerance, compared with approximately 89% for the proposed method. Across the evaluated range, our method maintains the highest cumulative pixel ratio, indicating more predictions with lower absolute depth errors in highly reflective regions.

3.2.2. Qualitative Analysis

To further evaluate the geometric quality of the reconstructed depth fields, we conduct a qualitative comparison in 3D space. Although the quantitative metrics provide an overall assessment of metric depth accuracy, they cannot fully characterize local geometric artifacts such as non-physical spikes, surface fragmentation, and topological collapse. These distortions are particularly relevant in highly reflective metal regions, where strong specular reflections may produce severe local errors. Therefore, it is also necessary to examine whether the predicted depth preserves physically plausible surface continuity and local structural consistency under severe photometric degradation.

Table 1. Quantitative comparison of monocular depth estimation on the XYZ-IBD dataset.

Method	Full Scene		Metal Region		Metal Region δ_1 ($\uparrow$)
	AbsRel ($\downarrow$)	RMSE ($\downarrow$)	AbsRel ($\downarrow$)	RMSE ($\downarrow$)	
DA-V2 [18]	0.1195	125.8 mm	0.1394	146.8 mm	0.4697
ZoeDepth [19]	0.1342	117.8 mm	0.1430	128.3 mm	0.8196
UniDepthV2 [7]	0.1027	99.7 mm	0.1158	111.4 mm	0.8498
Ours	0.0860	87.6 mm	0.0948	97.0 mm	0.9139

**Fig. 3.** Error Cumulative Distribution Function (CDF) Curves.

For this purpose, the predicted depth maps are back-projected into 3D space using the camera intrinsics provided with the test set. Figure 4 shows that the baseline models, including DA-V2, ZoeDepth, and UniDepthV2, exhibit non-physical spikes, surface fragmentation, and local topological distortion in highly reflective metal regions. These artifacts are especially evident around specular highlights and object boundaries, where unreliable photometric cues cause abrupt depth fluctuations and discontinuous surfaces. In contrast, the proposed method substantially suppresses these reflection-induced geometric artifacts and reconstructs more continuous and physically plausible 3D surfaces, providing a more reliable representation for subsequent point-cloud registration and grasp pose estimation.

Three main observations can be drawn from the visualization:

- **Physical Blind Spots of Active Sensors (GT):** Due to specular reflection and radiometric saturation, active optical sensors may suffer from severe depth voids in highly reflective metal regions.
- **Topological Distortions of Foundation Models:** Existing zero-shot models (DA-V2, ZoeDepth, UniDepthV2) exhibit substantial geometric degradation under intense photometric degradation. They generate pronounced topological distortions, non-physical “stalactite” spikes, and severe high-frequency

Table 2. Ablation Study of the Geometric-Radiometric Collaborative Framework

Model	(s, t)	$E_{\text{grad}} + E_{\text{smooth}}$	Metal Region	
			RMSE ($\downarrow$)	LSS ($\downarrow$)
M1	$\times$	$\times$	111.4 mm	5.32
M2	$\checkmark$	$\times$	98.3 mm	3.51
Ours	$\checkmark$	$\checkmark$	97.0 mm	3.04

wavy artifacts, which can reduce the reliability of downstream surface-normal and edge extraction.

- **Geometric Fidelity of Ours:** By introducing anti-hole gradient constraints and Radiometric Edge-Aware Regularization, our Geometric-Radiometric Collaborative Framework not only suppresses the prevalent high-frequency distortions and spikes observed in the baseline models but also reconstructs continuous, physically plausible surfaces within the GT blind spots.

Because lower numerical error does not necessarily imply higher geometric fidelity, this qualitative analysis demonstrates that our method reconstructs high-quality 3D manifolds under reflective conditions. This geometric coherence reduces the risk of feature mismatches, providing a reliable 3D baseline for rigorous vision-only registration and grasp pose estimation.

3.3. Ablation Studies

To clarify the role of each key module in the Geometric-Radiometric Collaborative Framework, we conduct ablation experiments on several representative configurations. Experimental results are shown in Table 2.

Based on the quantitative results in Tables 1 and 2, a key conclusion can be drawn: Although purely data-driven fine-tuning (M2) improves performance over the zero-shot baseline, the RMSE in the metal region remains 98.3 mm, and the LSS remains higher than that of the complete model at 3.51. This result suggests a trade-off between pointwise numerical fitting and geometric regularization in strongly

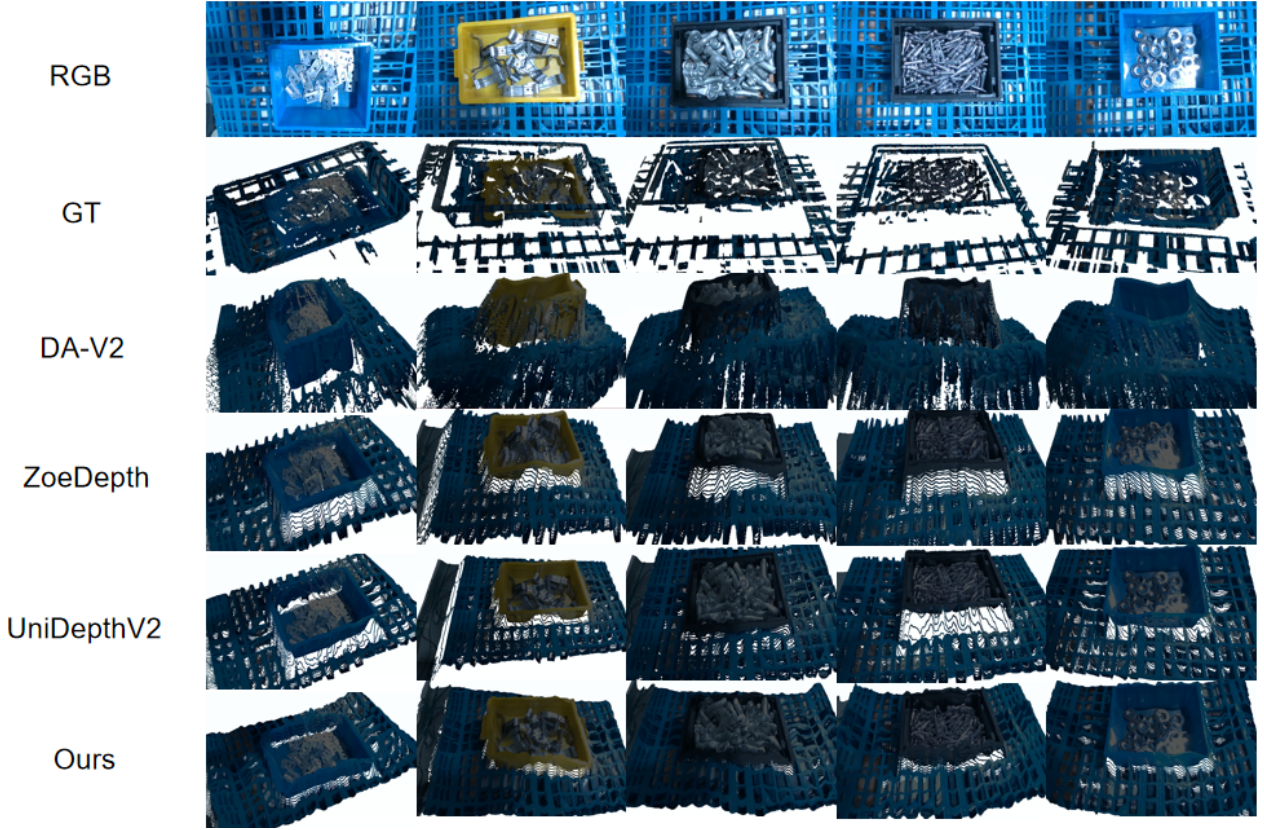


Fig. 4. Qualitative comparison of 3D point cloud reconstruction.

reflective regions. In industrial bin-picking scenarios, high-precision active sensors, which serve as the ground truth, inherently suffer from specular reflection effects, resulting in degraded, fragmented depth values and faint scanning noise along the edges of highly reflective metals. When a purely data-driven fine-tuning strategy relies solely on absolute error minimization, the network may also fit sensor noise near specular boundaries. From a spatial-manifold perspective, direct fitting to corrupted measurements can introduce large local gradient variations. Consequently, non-physical high-frequency spike artifacts may appear in the reconstructed 3D geometry, which can also limit further improvements in numerical accuracy.

These results indicate the importance of structural regularization. In contrast, the Geometric-Radiometric Collaborative Framework proposed in this paper, by introducing anti-hole gradient constraints (E_{grad}) and Radiometric Edge-Aware Regularization (E_{smooth}), not only effectively suppresses the non-physical spikes (LSS significantly reduced to 3.04) generated by the fine-tuning scheme (M2) in specular blind zones, but also further reduces metal-region RMSE to 97.0 mm by correcting large distortion errors. By abandoning blind numerical fitting in favor of adaptive

geometric interpolation, these results indicate that while preserving global metric accuracy, this paper’s framework improves local geometric fidelity. This high level of coherence at the geometric layer is beneficial for downstream robotic tasks reliant on surface normals for collision detection and grasp plane planning.

4. Conclusions

The proposed Geometric-Radiometric Collaborative Framework achieves an RMSE of 97.0 mm and a δ_1 accuracy of 91.39% in the metal region, while reducing the LSS to 3.04. These results indicate improvements in both metric depth accuracy and local surface smoothness under highly reflective conditions. This improved geometric fidelity provides a more reliable geometric basis for reducing collision-detection errors involving grippers or suction cups in downstream 6DoF grasp planning. Furthermore, the proposed intrinsic-free, single-shot calibration mechanism decouples metric scale recovery from complex multi-view geometry. This lightweight deployment paradigm avoids strict reliance on precise camera intrinsics, indicating potential robustness and practical value for real-world robotic bin-picking applications.

Although the current experiments primarily focus on 3D reconstruction accuracy at the vision-only level, we have yet to conduct large-scale, end-to-end disordered picking tests within a real closed-loop robotic-arm control system. In dynamic robotic environments, illumination changes, viewpoint variations, and occlusions may affect depth stability, while the additional energy-based refinement may impose computational overhead on real-time closed-loop control. To address this, future work will integrate this monocular metric depth estimation network into a practical robotic arm workstation (e.g., an eye-in-hand configuration). Combined with plane-grasp planning algorithms based on point-cloud normal vectors [20], future experiments will quantitatively evaluate the effects of the proposed method on collision avoidance and picking success rates in highly reflective, unstructured bin-picking scenarios.

Acknowledgments

This research was funded by the Fundamental Scientific Research Expenses for the Universities of Liaoning Province, grant numbers LJ212410144022, 2025YBXM0140, JG26DB390 and LJ232410144074, and project number 2025-MSLH-589.

References

- [1] A. Cordeiro, L. F. Rocha, J. Boaventura-Cunha, D. Figueiredo, and J. P. Souza, (2026) "A Review of Visual Perception for Robotic Bin-Picking" **Robotics and Autonomous Systems** 196: 105236. DOI: [10.1016/j.robot.2025.105236](https://doi.org/10.1016/j.robot.2025.105236).
- [2] R. L. Haugaard and A. G. Buch. "SurfEmb: Dense and Continuous Correspondence Distributions for Object Pose Estimation with Learnt Surface Embeddings". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, 6739–6748. DOI: [10.1109/CVPR52688.2022.00663](https://doi.org/10.1109/CVPR52688.2022.00663).
- [3] J. Huang, J. Hu, P. K. Yu, S. Ilic, M. Sundermeyer, and B. Busam. "XYZ-IBD: Benchmarking Robust 6D Object Pose Estimation under Real-World Industrial Complexity". In: *European Conference on Computer Vision (ECCV)*. URL: <https://eccv.ecva.net/virtual/2026/poster/3426>. 2026.
- [4] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, (2022) "Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer" **IEEE Transactions on Pattern Analysis and Machine Intelligence** 44(3): 1623–1637. DOI: [10.1109/TPAMI.2020.3019967](https://doi.org/10.1109/TPAMI.2020.3019967).
- [5] A. Eftekhari, A. Sax, J. Malik, and A. Zamir. "OmniData: A Scalable Pipeline for Making Multi-Task Mid-Level Vision Datasets From 3D Scans". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, 10766–10776. DOI: [10.1109/ICCV48922.2021.01061](https://doi.org/10.1109/ICCV48922.2021.01061).
- [6] W. Yin, J. Zhang, O. Wang, S. Niklaus, L. Mai, S. Chen, and C. Shen. "Learning to Recover 3D Scene Shape from a Single Image". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021, 204–213. DOI: [10.1109/CVPR46437.2021.00027](https://doi.org/10.1109/CVPR46437.2021.00027).
- [7] L. Piccinelli, C. Sakaridis, Y.-H. Yang, M. Segu, S. Li, W. Abbeloos, and L. Van Gool, (2026) "UniDepthV2: Universal Monocular Metric Depth Estimation Made Simpler" **IEEE Transactions on Pattern Analysis and Machine Intelligence** 48(3): 2354–2367. DOI: [10.1109/TPAMI.2025.3628473](https://doi.org/10.1109/TPAMI.2025.3628473).
- [8] J. Yang, J. Yao, and S. L. Waslander. "Active Pose Refinement for Textureless Shiny Objects using the Structured Light Camera". In: *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, 7550–7557. DOI: [10.1109/IROS58592.2024.10801300](https://doi.org/10.1109/IROS58592.2024.10801300).
- [9] S. F. Bhat, I. Alhashim, and P. Wonka. "AdaBins: Depth Estimation Using Adaptive Bins". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021, 4009–4018. DOI: [10.1109/CVPR46437.2021.00400](https://doi.org/10.1109/CVPR46437.2021.00400).
- [10] A. Costanzino, P. Z. Ramirez, M. Poggi, F. Tosi, S. Mattoccia, and L. Di Stefano. "Learning Depth Estimation for Transparent and Mirror Surfaces". In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, 9210–9221. DOI: [10.1109/ICCV51070.2023.00848](https://doi.org/10.1109/ICCV51070.2023.00848).
- [11] W. Choi, K. Hwang, M. Choi, K. Han, W. Choi, M. Shin, and S. Im, (2025) "Intrinsic Image Decomposition for Robust Self-supervised Monocular Depth Estimation on Reflective Surfaces" **Proceedings of the AAAI Conference on Artificial Intelligence** 39(3): 2555–2563. DOI: [10.1609/aaai.v39i3.32258](https://doi.org/10.1609/aaai.v39i3.32258).
- [12] K. He, J. Sun, and X. Tang, (2013) "Guided Image Filtering" **IEEE Transactions on Pattern Analysis and Machine Intelligence** 35(6): 1397–1409. DOI: [10.1109/TPAMI.2012.213](https://doi.org/10.1109/TPAMI.2012.213).

- [13] F. Liu, C. Shen, and G. Lin. "Deep Convolutional Neural Fields for Depth Estimation from a Single Image". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, 5162–5170. DOI: [10.1109/CVPR.2015.7299152](https://doi.org/10.1109/CVPR.2015.7299152).
- [14] S. A. Shafer, (1985) "Using Color to Separate Reflection Components" **Color Research & Application** 10(4): 210–218. DOI: [10.1002/col.5080100409](https://doi.org/10.1002/col.5080100409).
- [15] J. M. Facil, B. Ummenhofer, H. Zhou, L. Montesano, T. Brox, and J. Civera. "CAM-Convs: Camera-Aware Multi-Scale Convolutions for Single-View Depth". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, 11818–11827. DOI: [10.1109/CVPR.2019.01210](https://doi.org/10.1109/CVPR.2019.01210).
- [16] X. Cheng, P. Wang, and R. Yang. "Depth Estimation via Affinity Learned with Convolutional Spatial Propagation Network". In: *European Conference on Computer Vision (ECCV)*. Springer International Publishing, 2018, 108–125. DOI: [10.1007/978-3-030-01270-0_7](https://doi.org/10.1007/978-3-030-01270-0_7).
- [17] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim. "Visual Prompt Tuning". In: *European Conference on Computer Vision (ECCV)*. Springer Nature Switzerland, 2022, 709–727. DOI: [10.1007/978-3-031-19827-4_41](https://doi.org/10.1007/978-3-031-19827-4_41).
- [18] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao. "Depth Anything V2". In: *Advances in Neural Information Processing Systems*. 37. Curran Associates, Inc., 2024, 21875–21911. DOI: [10.52202/079017-0688](https://doi.org/10.52202/079017-0688).
- [19] S. F. Bhat, R. Birkel, D. Wofk, P. Wonka, and M. Müller, (2023) "ZoeDepth: Zero-shot Transfer by Combining Relative and Metric Depth" **arXiv preprint arXiv:2302.12288**: DOI: [10.48550/ARXIV.2302.12288](https://doi.org/10.48550/ARXIV.2302.12288).
- [20] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg. "Dex-Net 2.0: Deep Learning to Plan Robust Grasps with Synthetic Point Clouds and Analytic Grasp Metrics". In: *Robotics: Science and Systems (RSS)*. 2017. DOI: [10.15607/rss.2017.xiii.058](https://doi.org/10.15607/rss.2017.xiii.058).