machines

MDPI

Article

RG-RGD: Task-Triggered Small-Target RGB-D Depth Refinement for Robotic Laser Ablation

Bowen Si, Dayong Ning *, Jiaoyi Hou *, Yongjun Gong, Ming Yi, Fengrui Zhang and Zhilei Liu

Naval Architecture and Ocean Engineering College, Dalian Maritime University, Dalian 116026, China; sibw@dlmu.edu.cn (B.S.)

* Correspondence: ningdayong@dlmu.edu.cn (D.N.); houjiaoyi@vip.163.com (J.H.)

Abstract

Robotic 3D manipulation of small targets, such as laser ablation of urban-tree fruit balls, requires locally reliable depth. Mainstream RGB-D depth-refinement methods usually optimize image-wide metrics, which can leave task-relevant regions under-resolved and limit their direct use in precision robotic operation. This paper presents residual-gated RGB-D depth refinement (RG-RGD), a task-triggered local depth-refinement framework that reallocates computation toward task-relevant 3D geometry after a candidate target has been selected. The method contains three coupled designs: a self-play benefit-driven foveation mechanism that focuses network capacity on regions where refinement reduces geometric residuals; residual prediction with Bayesian measurement fusion that anchors predictions to available raw observations; and inertial measurement unit (IMU)-conditioned self-supervised training that improves inter-frame view consistency. Evaluated on the Visual Odometry with Inertial and Depth (VOID) benchmark, RG-RGD obtains competitive standard depth-completion metrics, including a mean absolute error (MAE) of 24.95 mm and an inverse mean absolute error (iMAE) of 10.85. On self-collected London plane fruit-ball sequences, the method reduces region-of-interest geometric error more strongly than full-image error. A qualitative demonstrative laser-ablation use case illustrates how refined local geometry can drive physical branch filtering, cutting-point selection, and gimbal-based execution. The results support task-triggered local depth refinement as a practical perception component for robotic manipulation.

Keywords: robotic laser ablation; RGB-D perception; depth refinement; small-target localization; self-supervised learning; foveated attention; visual servoing

Check for updates

Academic Editor: Bing Guo

Received: 30 June 2026
Revised: 22 July 2026
Accepted: 24 July 2026
Published: 25 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Robotic manipulation in natural or semi-structured environments often depends on a small part of the image rather than on the whole scene. For many robotic operations, the decisive quantity is not the average quality of a dense depth map, but whether the robot has reliable local 3D geometry at the pixels that determine the next physical action. A depth error near a contact point, cutting point, grasping point, or supporting branch can move a back-projected 3D target, change the tool-to-target distance, or make a controller select the wrong structure. Errors in distant background regions are less important if they are never consumed by the downstream robot. This difference makes local depth refinement a task-level robotic perception problem.

London plane fruit-ball laser ablation is used in this paper as a representative case. London plane trees are common urban trees, but mature fruit balls can release allergenic

trichomes and pollen in spring, motivating physical removal in some urban-management scenarios [1,2]. Manual pruning with aerial work vehicles is labor intensive and disruptive, while robotic pruning and horticultural manipulation remain limited by perception, localization, and end-effector constraints [3–6]. A laser end effector is mechanically light and non-contact, but it imposes a strict requirement on local depth: the branch connected to the selected fruit ball must be screened from background structures, the cutting point must be back-projected to a metric 3D coordinate, and the gimbal must be driven before the operation decision becomes stale.

RGB-D cameras are suitable for this setting because they provide metric depth in a compact sensing package. Their most useful measurements, however, are often least reliable near the structures that matter for manipulation. Structured-light and time-of-flight cameras can produce depth holes, flying pixels, multipath artifacts, and boundary degradation near thin branches, weakly textured fruit balls, and depth discontinuities [7,8]. Existing depth-completion networks have made clear progress on RGB-guided dense depth estimation [9–20]. Many of these methods are optimized for full-image metrics, road-scene LiDAR benchmarks, or globally supervised dense prediction. Such benchmarks remain necessary, but they do not test whether the largest errors occur at the small set of pixels consumed by a manipulation decision. Self-supervised depth learning and visual-inertial methods reduce the need for dense ground truth [21–28], but global photometric objectives can still be dominated by large background regions.

The central question addressed here is how to refine RGB-D depth when the downstream robot needs locally reliable 3D geometry rather than uniformly dense reconstruction. The proposed RG-RGD framework follows three principles. First, refinement capacity should be concentrated where it reduces task-relevant geometric residuals. Second, valid RGB-D measurements should remain metric anchors unless the network has evidence that correction is beneficial. Third, self-supervised training should use physical motion cues when dense ground-truth depth is unavailable. The public VOID dataset and a London plane laser-ablation use case are used as complementary evidence: VOID evaluates the transferable depth-refinement architecture under standard metrics, while the London plane data and robotic use case show how local geometry is consumed by downstream physical rules.

The main contributions are as follows:

1. A local RGB-D depth-refinement formulation is developed for robotic 3D manipulation, with the London plane laser-ablation pipeline serving as a representative small-target use case.
2. A self-play benefit-driven foveated scale head with small-object focused attention (BFS-SOFA) is introduced to resolve the circular dependency between the foveation mask and the focused prediction pass without manual region labels.
3. A residual-gated Bayesian measurement fusion and localized edge-aware propagation scheme is developed to preserve metric anchors, suppress cross-boundary diffusion, and focus correction on task-relevant regions.
4. An IMU-assisted self-supervised training pipeline is used to stabilize view-synthesis supervision when dense depth labels are unavailable in the robotic scene.
5. Evaluation combines VOID benchmark results, self-collected London plane ROI analysis, latency profiling, module-level timing, and a demonstrative laser-ablation use case.

2. Materials and Methods

The notation used in the following formulation is summarized in Table 1.

Table 1. Notation definitions.

Symbol	Definition
I_t	Input RGB image at time t
D_t^{raw}	Raw, potentially sparse or noisy depth observation from the RGB-D sensor
D_t^h	Dense depth hint generated from valid raw measurements
ΔD_t	Residual predicted around the dense hint
$\hat{D}_t$	Non-negative network prediction before measurement fusion
D_t^{bmf}	Depth map after Bayesian measurement fusion
D_t^o	Final refined local depth map after propagation
M_t	Binary validity mask of the raw depth observation
σ_t^2, $\sigma_{obs,t}^2$	Predicted network and observation variances used by BMF
α_t	BMF coefficient assigned to a valid raw observation
F_t, z_t	Fused RGB-D features and benefit-head logits
τ_t, T	Image-adaptive threshold and fixed sigmoid temperature
B_t, $\widetilde{B}_t$	Raw and normalized self-play benefit targets
m_{so}	Task-relevant foveation mask generated by the benefit field
R_t	Task-relevant image region used by the next robotic decision
o_t	Boolean mask of valid, positive raw-depth observations
$\mathcal{P}_9(p)$	Center pixel and its eight local propagation neighbors
u_{pq}, w_{pq}, g_p	Affinity logit, edge-normalized affinity, and ROI-modulated propagation gate
K	Number of UACSPN propagation iterations
$\mathbf{u}_{ts}$, $\mathbf{z}_{ts}$	Synchronized 12-D IMU feature and its normalized, clipped form
$\mathbf{p}_{ts}^v$, $\mathbf{p}_{ts}$	Vision-only and IMU-conditioned six-dimensional pose logits

2.1. Problem Formulation and System Overview

Reliable local 3D geometry is the quantity required by the downstream robot. For a selected manipulation target, let R_t denote the task-relevant region at time t, including the target, nearby supporting structures, depth boundaries, and pixels used by the next decision rule. Given an aligned RGB image I_t and raw depth D_t^{raw}, the objective is to produce a refined depth map D_t^o that reduces metric error inside R_t:

$$\min_{\theta} \ \mathcal{E}_{geo}(D_\theta, R_t) \tag{1}$$

where θ denotes model parameters.

The architecture therefore avoids treating all pixels as equally important. It first constructs a dense depth hint from valid measurements:

$$D_t^h = \mathcal{N}(D_t^{raw}, M_t), \tag{2}$$

where $\mathcal{N}(\cdot)$ denotes nearest-neighbor propagation over valid depth samples. Hybrid RGB-D features are then extracted and fused:

$$F_t = \Phi\Big(E_I(I_t), E_D(D_t^h, D_t^{raw}, M_t)\Big), \tag{3}$$

where E_I and E_D are the image and depth encoders and Φ is cross-modal fusion. The fused features are passed to a benefit-driven foveation module, a residual-gated Bayesian measurement fusion module, and an edge-aware local propagation module. Figure 1 summarizes the complete pipeline.

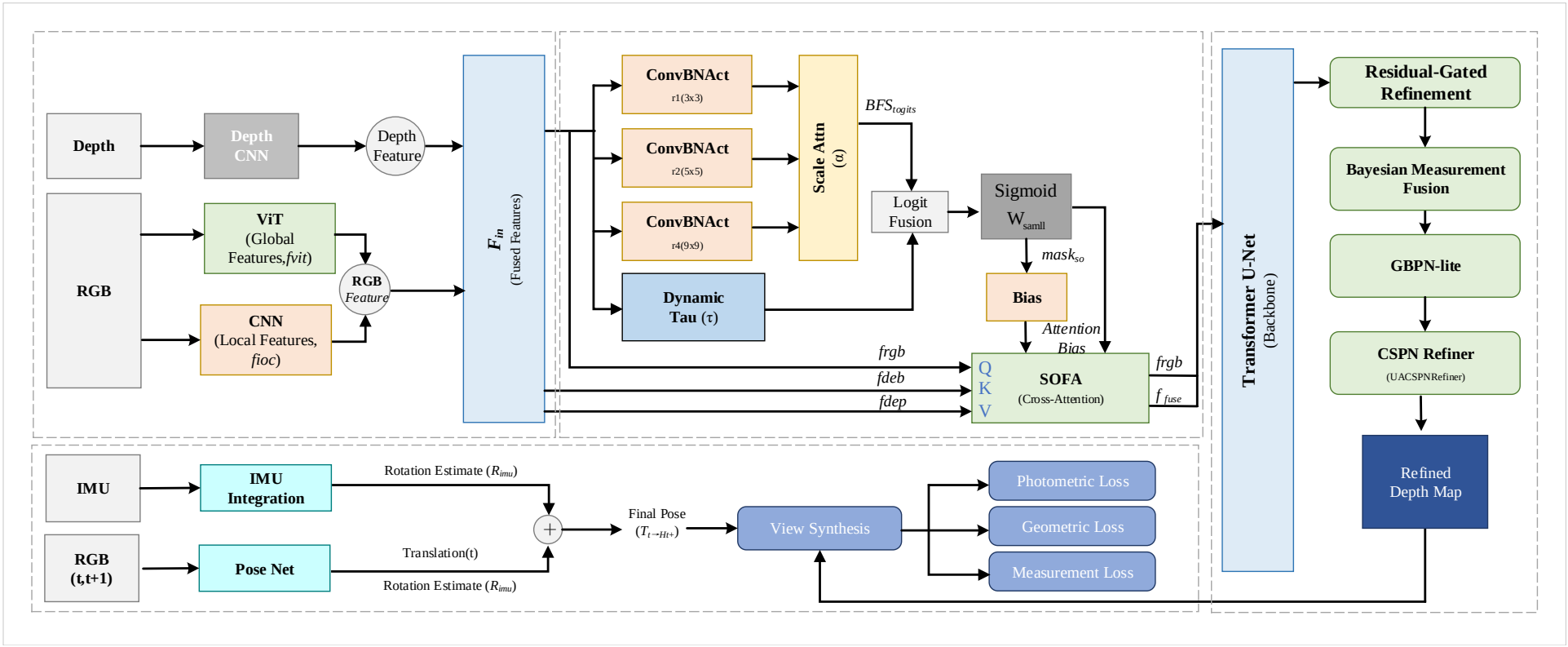


Figure 1. Overall architecture of RG-RGD. RGB and depth features are fused before benefit-driven foveation and small-object focused attention (SOFA) cross-attention. The resulting features enter the Transformer U-Net backbone, followed by residual prediction, Bayesian measurement fusion (BMF), Gaussian belief-propagation network (GBPN)-lite, and uncertainty-aware convex spatial propagation network (UACSPN) refinement. The lower branch provides self-supervised training constraints.

https://doi.org/10.3390/machines14080841

The nearest-neighbor operator is used only to construct a dense conditioning hint. It can copy a wrong value across a depth boundary when raw samples are sparse or contain flying pixels. The original D_t^{raw} and M_t are therefore retained as separate network inputs, and the interpolated hint is not treated as a reliable sensor observation.

The architecture follows three constraints from robotic 3D operation. First, task-relevant structures occupy only a small fraction of the image, so refinement should concentrate on high-benefit regions. Second, raw RGB-D measurements provide metric scale and should be preserved when reliable. Third, the method must remain compatible with an online system, so the foveation module is separable from the shared RGB-D feature extractor.

2.2. Benefit-Driven Foveation via Self-Play Supervision

Benefit-driven foveated scale head with small-object focused attention (BFS-SOFA) is a lightweight resource-allocation module. It is related to foveated visual attention, spatial transformation, and dynamic token selection [29–35], but its focusing signal is defined by depth-refinement benefit rather than visual saliency. A direct formulation creates a circular dependency: the foveation mask should be learned from the error distribution, but the error distribution changes after foveation is applied.

The full network combines lightweight convolutional operations [36] with an encoder–decoder topology based on U-Net and Transformer U-Net designs [37,38].

BFS-SOFA forms three dilated feature branches, r_1, r_2, r_4, and their per-pixel softmax scale weights, a_1, a_2, a_4. The benefit-head logit and image-adaptive threshold are

$$z_t(p) = \sum_{s\in\{1,2,4\}} a_s(p)\, h_s(r_s)(p), \qquad \tau_t = h_\tau(\mathrm{GAP}(F_t)), \tag{4}$$

where h_s and h_τ are learned projections and GAP is global average pooling. A small-object proxy is added as a logit bias and the resulting map is optionally Gaussian-smoothed before thresholding. With the default temperature $T = 0.1$, the mask used by SOFA is

$$m_{so}(p) = \mathrm{clip}_{[0,1]}\left[\sigma\left(\frac{\mathcal{G}(z_t(p) + b_{proxy}(p)) - \tau_t}{T}\right)(a_1(p) + a_2(p))\frac{p_0}{\overline{m} + \epsilon}\right], \tag{5}$$

where $\mathcal{G}$ denotes the optional 7×7 Gaussian smoothing used in the default configuration, $p_0 = 0.05$ is the target mask mass, and $\overline{m}$ is the image-wise mean before budget normalization. The last factor is applied only when budget normalization is enabled. Thus, τ_t is dynamic for every image, whereas T controls the sharpness of the sigmoid.

To break the circular dependency, training uses a baseline pass without BFS-SOFA and a focused pass with BFS-SOFA. The reduction in measurement and photometric residuals defines the detached benefit target:

$$B_t(p) = \mathrm{ReLU}\Big(e^m_{base}(p) - e^m_{fov}(p)\Big)M_t(p) + \eta\ \mathrm{ReLU}\Big(e^{ph}_{base}(p) - e^{ph}_{fov}(p)\Big)M_{ph}(p), \tag{6}$$

The normalized target and benefit loss are

$$\widetilde{B}_t = \begin{cases} \mathrm{clip}_{[0,1]}(B_t/\beta), & \beta > 0, \\ \mathrm{clip}_{[0,1]}\big(B_t/(Q_q(B_t) + \epsilon)\big), & \beta \le 0, \end{cases} \qquad \mathcal{L}_{benefit} = \frac{1}{|\Omega|}\sum_{p\in\Omega} \mathrm{BCEWithLogits}\Big(z_t(p), \widetilde{B}_t(p)\Big), \tag{7}$$

where M_{ph} denotes valid reprojection pixels, Ω is the valid support, and Q_q is its benefit quantile. The reported configuration uses $\eta = 0.5$, $\beta = 0.15$, $q = 0.90$ for self-supervised training ($q = 0.95$ on VOID), and $\lambda_b = 0.1$.

This two-pass mechanism makes the benefit map a training signal derived from actual improvement, while the stop-gradient operation prevents the baseline pass from

being optimized to make the focused pass appear beneficial. Figure 2 summarizes the benefit-driven foveated attention module.

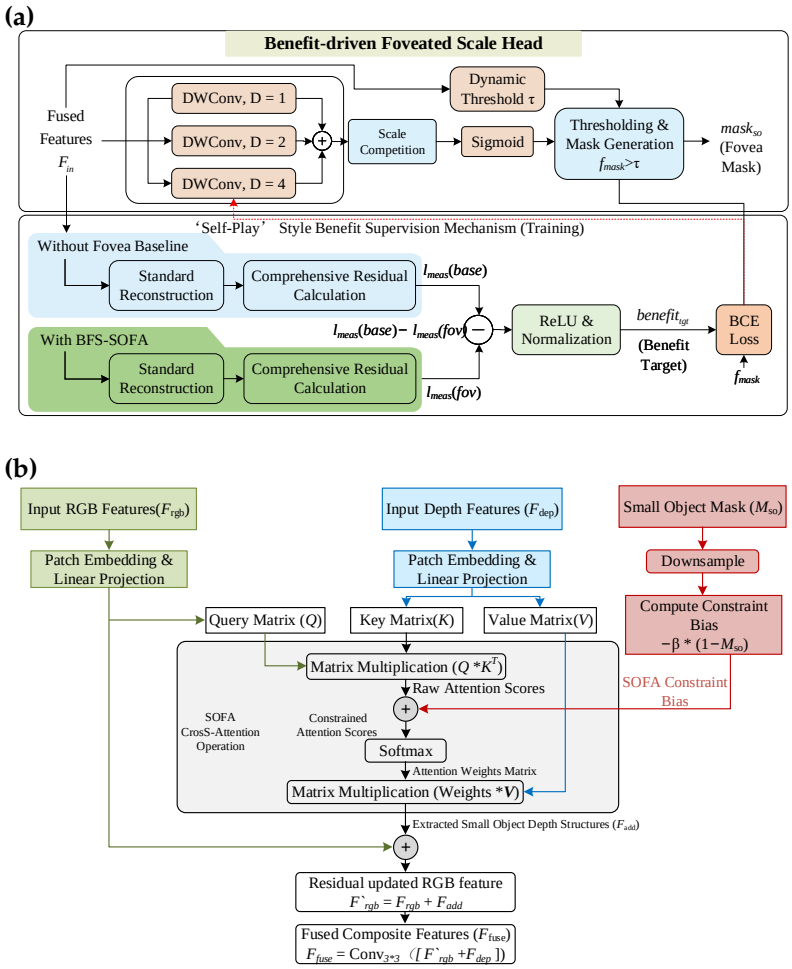


Figure 2. Benefit-driven foveated attention. (**a**) The foveated scale head combines three receptive-field scales and uses an image-adaptive threshold to produce the normalized mask $m_{so} \in [0, 1]$. (**b**) Small-object focused attention uses this mask to bias feature interaction toward thin structures, object boundaries, and depth discontinuities. In panel (**b**), the asterisks denote multiplication, and the minus sign in $1 - M_{so}$ denotes subtraction.

2.3. Residual-Gated Bayesian Measurement Fusion

Directly predicting absolute depth can overwrite valid sensor measurements. RG-RGD therefore predicts a residual around the dense depth hint:

$$\hat{D}_t = \mathrm{clip}_{[0,D_{max}]}\Big(\mathrm{ReLU}\Big(D_t^h + \Delta D_t\Big)\Big), \tag{8}$$

where $D_{max} = 10$ m.

The network also predicts positive standard deviations for the depth estimate and the sensor observation:

$$\sigma_t = \mathrm{clip}(\mathrm{softplus}(h_\sigma(F_t)) + \epsilon,\, 0.02, 5.0), \quad \sigma_{obs,t} = \mathrm{clip}(\mathrm{softplus}(h_{obs}(F_t, m_{so}, M_t)) + \epsilon,\, 0.01, 5.0). \tag{9}$$

The corresponding variances are $v_t = \sigma_t^2$ and $v_{obs,t} = \sigma_{obs,t}^2$. The observation variance is learned from valid raw-depth pixels using a Gaussian measurement likelihood, while the fixed bounds prevent degenerate values.

For $o_t = M_t \wedge (D_t^{raw} > \epsilon)$, the fusion coefficient and refined estimate are

$$\alpha_t = o_t \frac{v_t}{v_t + v_{obs,t}}, \qquad D_t^{bmf} = \alpha_t D_t^{raw} + (1 - \alpha_t)\hat{D}_t. \tag{10}$$

At unobserved pixels, $\alpha_t = 0$ and $\hat{D}_t$ is retained. At observed pixels, α_t increases with prediction uncertainty and decreases with observation uncertainty, preserving reliable metric measurements while allowing local correction.

2.4. Edge-Aware Propagation with UACSPN

Spatial propagation further removes isolated artifacts after BMF, but propagation can blur across object boundaries if affinities are learned only from depth features. This is undesirable for small-target manipulation because centimeter-scale discontinuities can determine whether a branch or background structure is selected. RG-RGD therefore uses an uncertainty-aware convex spatial propagation network (UACSPN) with three explicit constraints: an RGB edge barrier, ROI-boosted gating, and observation anchoring.

UACSPN uses a nine-position support $\mathcal{P}_9(p)$ consisting of the center and its eight neighbors. It starts from the preceding learned local refiner, $D_t^{(0)} = \mathrm{BP}(D_t^{bmf})$. A small convolutional head predicts affinity logits u_{pq} and a gate logit. The initial affinities are softmax-normalized, then an RGB-gradient edge barrier is applied and the affinities are renormalized:

$$\bar{w}_{pq} = \mathrm{softmax}_{q \in \mathcal{P}_9(p)}(u_{pq}), \qquad w_{pq} = \frac{\bar{w}_{pq} \exp(-\,\mathrm{softplus}(a)\, e_{pq})}{\sum_{r \in \mathcal{P}_9(p)} \bar{w}_{pr} \exp(-\,\mathrm{softplus}(a)\, e_{pr}) + \epsilon}, \tag{11}$$

where e_{pq} is the horizontal, vertical, or averaged diagonal RGB gradient between the two positions and a is a learned scalar. The center affinity has barrier one. The propagation gate is

$$g_p = \sigma(g_p^{raw})(0.35 + 0.65 m_{so}(p)), \qquad \widetilde{D}_t^{(k+1)}(p) = D_t^{(k)}(p) + g_p \left(\sum_{q \in \mathcal{P}_9(p)} w_{pq} D_t^{(k)}(q) - D_t^{(k)}(p) \right). \tag{12}$$

After every update, observed pixels are re-anchored to the detached BMF output:

$$D_t^{(k+1)}(p) = \begin{cases} \mathrm{StopGrad}(D_t^{bmf}(p)), & o_t(p) = 1, \\ \widetilde{D}_t^{(k+1)}(p), & o_t(p) = 0. \end{cases} \tag{13}$$

Four propagation iterations are used ($K = 4$). Re-anchoring preserves fused metric observations while allowing surrounding holes and uncertain regions to be refined. After the final iteration, $D_t^o = D_t^{(K)}$. Table 2 summarizes the conceptual differences between NLSPN and UACSPN.

Table 2. Conceptual differences between NLSPN and the local UACSPN implementation used in RG-RGD.

Aspect	NLSPN [12]	UACSPN in RG-RGD
Neighborhood	Non-local learned offsets	Local center plus 8-connected neighborhood
Affinity evidence	Learned affinity from features	RGB edge barrier with base affinity
Propagation gate	Per-pixel learned confidence	ROI-boosted gate from m_{so}
Measurement anchoring	No hard re-injection per step	Validity-mask re-injection after each step
Failure mode addressed	General depth completion	Cross-boundary bleeding in local robotic regions

2.5. IMU-Assisted Self-Supervised Training Pipeline

Because dense ground-truth depth is unavailable for the self-collected sequences, training uses temporally adjacent RGB-D frames. The predicted depth and relative pose are used to warp a neighboring RGB image into the current view, and reconstruction consistency provides the self-supervised signal. The relative pose between two adjacent frames is represented as

$$T_{t\to s} = \begin{bmatrix} R_{t\to s} & \mathbf{t}_{t\to s} \\ \mathbf{0}^\top & 1 \end{bmatrix}. \tag{14}$$

A lightweight PoseNet predicts six-dimensional pose logits $\mathbf{p}_{ts}^{v}$ from the RGB pair. The synchronized IMU window is summarized by integrated angular increment, mean acceleration, mean angular velocity, and angular-velocity standard deviation, then normalized with training-partition statistics:

$$\mathbf{z}_{ts} = \mathrm{clip}_{[-5,5]}\left(\frac{\mathbf{u}_{ts} - \boldsymbol{\mu}_{tr}}{\boldsymbol{\sigma}_{tr} + \epsilon}\right). \tag{15}$$

An IMU residual adapter conditions the visual logits:

$$\mathbf{p}_{ts} = \mathbf{p}_{ts}^{v} + \sigma(\boldsymbol{\gamma}) \odot f_{imu}([\mathrm{StopGrad}(\mathbf{p}_{ts}^{v}), \mathbf{z}_{ts}]). \tag{16}$$

The conditioned logits are mapped to translation and axis–angle rotation to construct $T_{t\to s}$. With camera intrinsics K, a target-frame pixel is projected into a neighboring source frame as

$$p_s \sim K T_{t\to s} D_t^{o}(p_t) K^{-1} p_t. \tag{17}$$

Bilinear sampling reconstructs the target image from the source image. The training objective combines photometric reprojection, sparse measurement consistency, edge-aware smoothness, and benefit supervision:

$$\mathcal{L}_{total} = \lambda_p \mathcal{L}_{photo} + \lambda_s \mathcal{L}_{sparse} + \lambda_g \mathcal{L}_{smooth} + \lambda_b \mathcal{L}_{benefit}. \tag{18}$$

Training follows a two-stage schedule. The RGB-D encoder, residual head, BMF module, and UACSPN module are first optimized with photometric, sparse measurement, and smoothness losses. BFS-SOFA is then activated and trained with Algorithm 1. Figure 3 illustrates the training loop.

Algorithm 1: Self-play benefit supervision training step

Input: RGB image I_t, raw depth D_t^{raw}, current network parameters θ

Output: Updated network parameters θ

1. Disable BFS-SOFA for the baseline forward pass.
2. Predict $\hat{D}_{base} = \mathrm{Net}(I_t, D_t^{raw}; \theta)$.
3. Compute pixel-wise baseline loss ℓ_{base} and detach it from gradient flow.
4. Enable BFS-SOFA and predict $m_{so} = \mathrm{BFS}(I_t, D_t^{raw}; \theta)$.
5. Predict $\hat{D}_{fov} = \mathrm{Net}(I_t, D_t^{raw}, m_{so}; \theta)$.
6. Compute pixel-wise focused loss ℓ_{fov}.
7. Form the detached measurement benefit and, when enabled, the warp-valid photometric benefit.
8. Normalize B_t to $\widetilde{B}_t$ and compute BCEWithLogits between z_t and $\widetilde{B}_t$.
9. Update θ using the focused-pass reconstruction losses and $\lambda_b \mathcal{L}_{benefit}$.

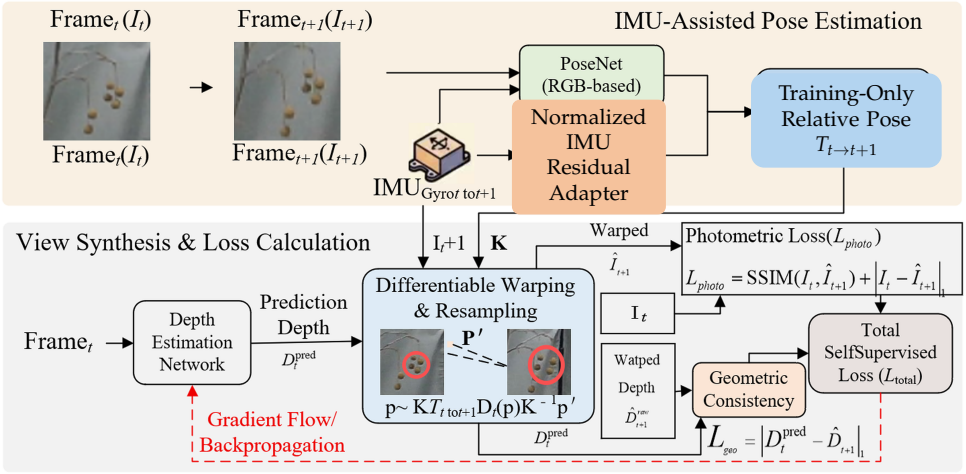


Figure 3. IMU-assisted self-supervised training loop. The RGB pair produces a visual pose estimate, and the normalized IMU residual conditions view-synthesis supervision. Raw-depth consistency anchors metric scale. The red circles mark corresponding target regions before and after differentiable warping.

3. Results

3.1. Experimental Data and Protocols

Experiments included evaluation on the Visual Odometry with Inertial and Depth (VOID) benchmark, self-collected robotic data, runtime profiling, and a qualitative robotic use case. VOID was used to compare RG-RGD with representative depth-completion methods under standard geometric metrics.

The self-collected data comprised three indoor RGB-D and IMU sequences of London plane fruit balls with 220, 239, and 293 synchronized frames (752 frames in total). A Gemini 335L RGB-D camera (Orbbec Technology Group Co., Ltd., Shenzhen, China) recorded 1280 × 720 RGB and aligned depth at 30 Hz, and the gyroscope stream had a median sampling rate of approximately 992 Hz. Raw depth was aligned to the color stream, and the stored RGB intrinsics and gyro-to-color rotation were used for calibration. The targets were approximately 2–3 cm in diameter and were captured at a working distance of about 2.9 m under thin-branch occlusion, weak texture, background clutter, and illumination variation.

The first two sequences (459 frames; 528 valid temporal pairs) were used for training, and the third sequence (293 frames; 218 pairs) was used for validation. Dense ground-truth depth was unavailable for these sequences; they were therefore used for self-supervised training, ROI analysis, timing evaluation, and the qualitative use case.

3.2. Depth-Completion Benchmark

Table 3 compares RG-RGD with representative methods on VOID. RG-RGD obtains the lowest MAE and iMAE among the listed methods, while its RMSE and iRMSE remain higher than those of CostDCNet and NLSPN. On the same 800-sample test split, the P95 and P99 absolute errors are 88.5 mm and 294.0 mm, respectively, indicating that large boundary and occlusion errors remain.

Figure 4 shows representative VOID examples. Each example contains RGB input, sparse/raw depth, RG-RGD refined depth, an error map, foveation response, and a reconstructed view. The foveation maps are strongest near small objects and depth discontinuities, consistent with the intended behavior of the benefit-driven module.

Table 3. Comparison on the VOID depth-completion benchmark. Lower values are better.

No.	Method	Venue/Year	MAE	RMSE	iMAE	iRMSE
1	RG-RGD	This work	24.95	94.99	10.85	44.52
2	CostDCNet	ECCV 2022	25.84	76.28	12.19	32.13
3	NLSPN	ECCV 2020	26.74	79.12	12.70	33.88
4	Monitored Dist.	ECCV 2022	29.67	79.78	14.84	37.88
5	PENet	ICRA 2021	34.61	82.01	18.89	40.36
6	MSG-CHN	WACV 2020	43.57	109.94	23.44	52.09
7	ENet	ICRA 2021	46.90	94.35	26.78	52.58
8	Scanline RI-DC	RA-L/IROS 2021	59.40	181.42	19.37	46.56

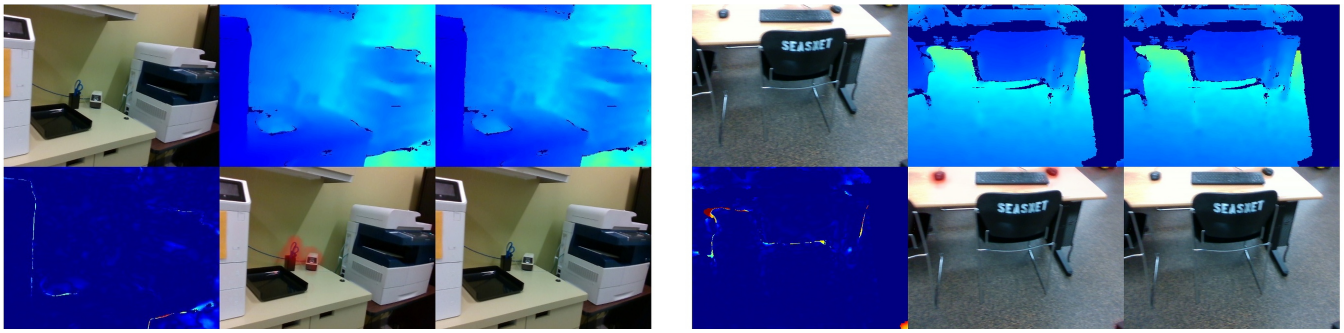


Figure 4. VOID qualitative depth-refinement examples. Two benchmark examples are placed side by side. In each example, the six tiles are arranged from left to right and top to bottom as RGB input, sparse/raw depth, RG-RGD refined depth, absolute-error visualization, foveation response, and reconstructed view. Black pixels in the raw-depth views are invalid observations. Depth values originate in meters; the false-color depth and error tiles are normalized separately for visualization and are not intended for cross-example numerical comparison. Quantitative depth errors are reported in millimeters in Table 3; the foveation response is normalized to $[0, 1]$.

3.3. Self-Collected Fruit-Ball Examples and Focused Ablation

Figure 5 shows eight indoor frames from the self-collected London plane fruit-ball sequence. Each row contains RGB input, raw depth, refined depth, and the BFS-SOFA overlay. The overlays concentrate on fruit balls, thin branches, and boundary regions, indicating that the module assigns refinement capacity to the structures used by branch screening and cutting-point selection.

Table 4 reports the focused ablation of BFS-SOFA on the self-collected validation set. Adding BFS-SOFA reduces the full-image geometric error from 0.0079 to 0.0073 and the region-of-interest (ROI) geometric error from 0.0090 to 0.0076. The larger ROI gain is important for the robotic task because branch localization depends on local boundary and thin-structure pixels. Photometric error and the structural similarity index measure (SSIM) remain nearly unchanged, suggesting that the main effect is local geometric improvement rather than a global image-similarity change.

Table 4. Ablation study of BFS-SOFA on the self-collected validation set.

Metric	Base	BFS-SOFA
Full-image geometric error	0.0079	0.0073
Relative improvement	–	7.5%
ROI geometric error	0.0090	0.0076
ROI relative improvement	–	15.3%
Full-image photometric error	0.0530	0.0530
Full-image SSIM	0.6921	0.6921

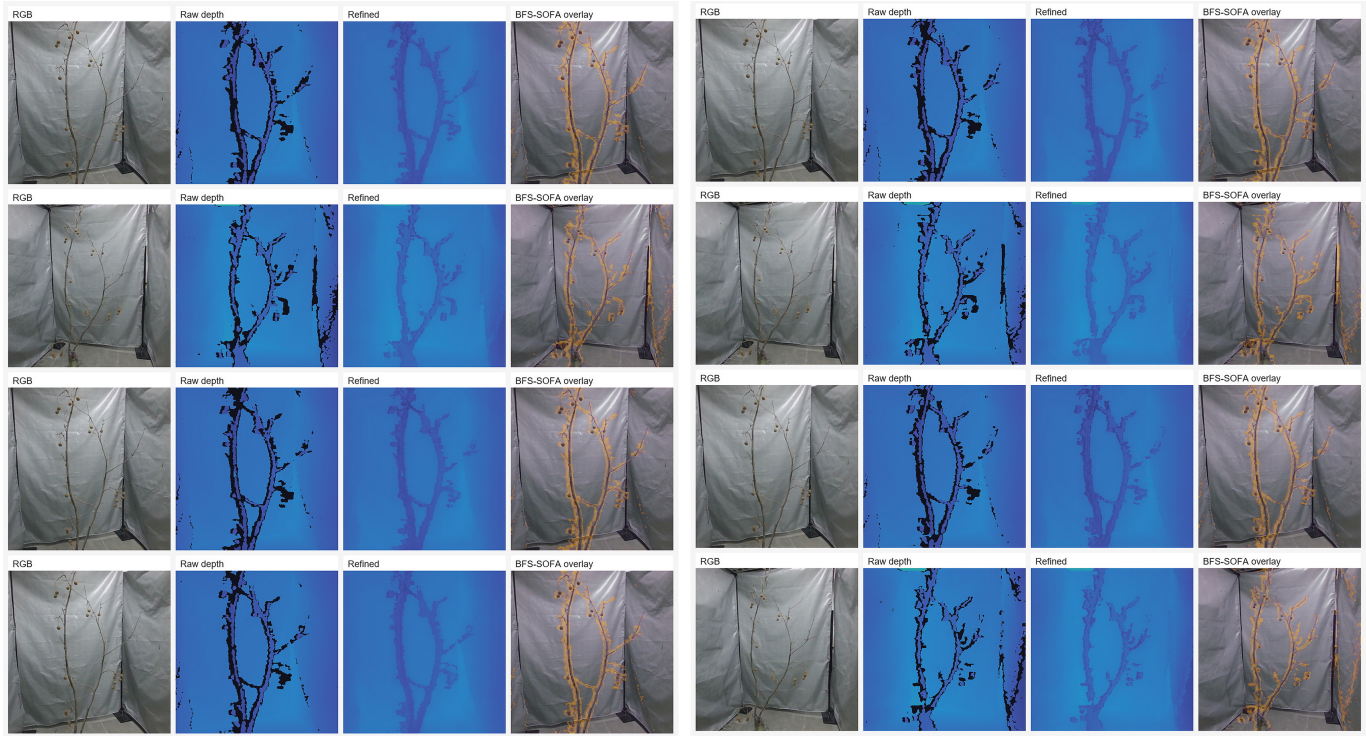


Figure 5. Indoor BFS-SOFA small-target focusing examples. Eight self-collected fruit-ball frames are shown in two blocks. In each row, the four tiles are RGB input, raw depth, RG-RGD refined depth, and BFS-SOFA overlay. Black regions in raw depth denote invalid observations; depth colors are display-normalized per frame and are not a common metric scale. The orange overlay is the normalized BFS-SOFA response $m_{s0} \in [0, 1]$, highlighting fruit balls, branch segments, and boundary regions used by the depth-refinement pipeline.

Table 5 compares direct residual replacement with the complete RG-BMF formulation on the VOID test split. BMF reduces MAE from 25.014 to 24.951 mm and iMAE from 10.892 to 10.853, with similar RMSE and tail-error values.

Table 5. BMF ablation on the 800-sample VOID test split. MAE, RMSE, P95, and P99 are in millimeters; iMAE and iRMSE are in km^{-1}. Bold values indicate the best result for each metric.

Variant	MAE	RMSE	iMAE	iRMSE	P95	P99
Direct replacement (BMF disabled)	25.014	94.994	10.892	44.523	88.5	294.0
Full RG-BMF	**24.951**	**94.992**	**10.853**	**44.518**	88.5	294.0

The IMU-assisted pose model was evaluated on the same 218-pair validation sequence. As shown in Table 6, IMU conditioning reduces the total, photometric, geometric, and warp consistency losses by 12.72%, 16.34%, 17.76%, and 20.03%, respectively.

Table 6. IMU ablation on the self-collected validation sequence (218 temporal pairs). Lower consistency loss is better. Bold values indicate the best result for each metric.

Pose-Conditioning Variant	Total	Photometric	Geometry	Warp
Vision-only PoseNet	0.491303	0.240267	0.065215	0.072837
Normalized gated IMU residual	**0.428792**	**0.201010**	**0.053635**	**0.058245**

3.4. Latency Breakdown and Module-Level Timing

Runtime was measured on 300 RGB-D frames at 320 × 320 resolution using an NVIDIA GeForce RTX 3070 Ti laptop GPU (NVIDIA Corporation, Santa Clara, CA, USA). Automatic

mixed precision was enabled, and 30 warm-up iterations were used before timing. The model contains 24.89M parameters, with a 94.96 MB checkpoint and 1163.69 MB peak GPU memory use in the tested configuration. Table 7 reports both model and end-to-end pipeline latency. The network reaches 44.57 ms mean latency and 49.45 ms P95 latency. The complete Python pipeline reaches 72.70 ms mean latency and 78.05 ms P95 latency, corresponding to 13.75 FPS. Preprocessing is the second-largest cost because the current implementation generates the depth hint online in Python 3.10.

Table 7. Latency breakdown of the task-triggered RG-RGD path. Timing used 300 RGB-D frames at 320 × 320 resolution on an RTX 3070 Ti laptop GPU.

Stage	Mean	Median	P95
Preprocessing	27.46 ms	27.39 ms	28.76 ms
Host-to-device transfer	0.50 ms	0.48 ms	0.64 ms
Model inference	44.57 ms	43.70 ms	49.45 ms
Post-processing	0.06 ms	0.06 ms	0.09 ms
End-to-end Python pipeline	72.70 ms	71.76 ms	78.05 ms

An additional module-level timing experiment was used to isolate the cost of BFS-SOFA. Inputs were preloaded on the GPU to exclude disk access, OpenCV preprocessing, and host-to-device transfer. Timing used CUDA events, 30 warm-up iterations, and 300 timed runs. The module-only setting includes the benefit-driven foveated scale head and SOFA but excludes the shared RGB-D feature extractor. As shown in Table 8, BFS-SOFA adds about 12 ms under this protocol.

Table 8. Module-level timing of BFS-SOFA at 320 × 320 resolution. Inputs were preloaded on the GPU; timing used CUDA events with 30 warm-up iterations and 300 timed runs.

Variant	Mean	Median	P95	ΔMean
Base without BFS-SOFA	29.81 ms	29.88 ms	31.83 ms	–
Base + BFS-SOFA	41.85 ms	41.39 ms	45.51 ms	+12.04 ms
BFS-SOFA module only	12.69 ms	12.47 ms	15.60 ms	+12.69 ms

The measured latency corresponds to the target-triggered refinement step between target selection and branch screening. Compiled depth-hint generation and embedded-GPU optimization could further reduce this cost.

3.5. Qualitative Demonstrative Use Case: Robotic Laser Ablation

The London plane fruit-ball system is used as a demonstrative use case for how local refined geometry can be consumed by a robotic pipeline. The system contains an RGB-D camera, a two-axis gimbal, and a laser head. It follows the broader direction of robotic horticulture systems that connect perception to manipulation or pruning decisions [39,40]. Figure 6 shows the indoor setup, the laser head, and an outdoor simulated canopy scene.

The purpose of this use case is to illustrate how the refined local depth map triggers downstream physical rules. After a fruit ball is selected, the refined depth is used to back-project the target neighborhood and screen candidate branch pixels by three gates. First, depth-threshold filtering removes pixels outside the valid working range of the laser head. Second, angle filtering favors branch segments whose local direction points toward the selected fruit-ball center. Third, connected-component filtering suppresses crossing branches that appear behind the target and re-enter on the other side, because these structures are not physically connected to the selected fruit ball in the depth field. Figure 7 shows three outdoor examples of this filtering process.

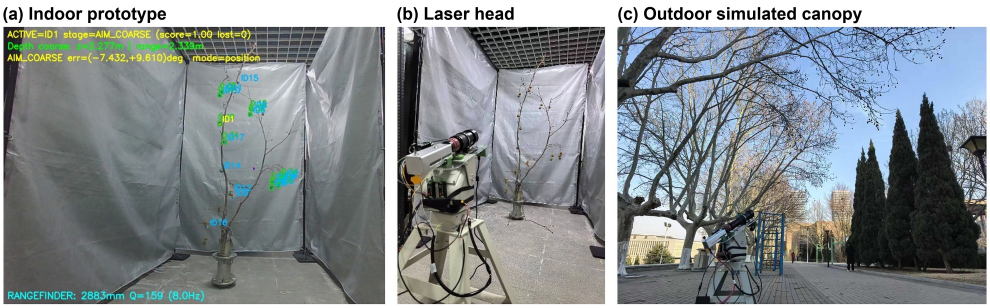


Figure 6. Qualitative demonstrative use case for London plane fruit-ball laser ablation. (**a**) Indoor prototype with an SK60 laser rangefinder (Shanghai Senky Photoelectric Technology Co., Ltd., Shanghai, China) showing 2883 mm for the illustrated capture; (**b**) two-axis gimbal and laser head; and (**c**) outdoor simulated-canopy context.



Figure 7. Qualitative examples of the cutting-point filtering rules. The left, middle, and right panels show three outdoor target cases. Cyan boxes mark the local target neighborhoods after depth-range pruning; small orange point markers indicate the selected fruit-ball and nearby branch pixels; magenta labels report the final camera-frame cutting-point coordinates $[X, Y, Z]$ in millimeters after angle and connected-component screening.

The retained branch segment then supplies a cutting-point candidate near the branch–fruit connection, which is transformed from the camera coordinate frame to the gimbal coordinate frame for laser pointing. Local RGB-D refinement can provide the metric geometry required by branch filtering and gimbal execution, but larger paired outdoor trials, confidence intervals, longer working distances, and stronger environmental disturbance tests are still needed before making quantitative performance claims.

4. Discussion

The results show why robotic local depth refinement should not be evaluated only as a full-image depth-completion problem. Benchmark metrics are useful for comparing geometric quality, but the robot depends on a smaller group of pixels around the target and nearby support structures. The focused ablation confirms this difference: BFS-SOFA gives a larger improvement in the ROI than in the full image. This behavior is valuable for physical filtering rules, where local boundary quality can determine whether a candidate structure is selected or rejected.

Measurement anchoring is also important. Self-supervised view synthesis can encourage visually plausible depth, but local photometric agreement does not guarantee metric correctness around thin structures. The residual prediction and BMF gate preserve a metric connection to available raw observations while allowing a correction in holes and uncertain regions. This design is consistent with the strong MAE and iMAE performance on VOID, while the weaker RMSE and iRMSE indicate that severe outliers are not fully eliminated.

The higher RMSE and iRMSE arise mainly from large errors at depth boundaries, connected holes, partial occlusions, flying pixels, and temporal reprojection mismatches.

The P95 and P99 results confirm this long error tail. BFS-SOFA reduces both full-image and ROI geometric error in the self-collected ablation, but isolated large errors remain.

The latency analysis separates model latency, preprocessing, tail latency, and module overhead. BFS-SOFA adds about 12 ms and remains separable from the shared RGB-D feature extractor, supporting its use as a portable add-on for other RGB-D refinement backbones.

The demonstrative use case also clarifies the main failure sources that motivate local refinement. Missing depth at fruit-ball boundaries and thin branch junctions can shift the back-projected cutting point, while crossing branches may interfere with connected-component screening. Before laser execution, the system reacquires RGB-D observations through the visual loop, and the two-axis gimbal uses an independent angle-control loop. Strong occlusion, reflection, wind, and branch motion can still affect localization.

Safety Considerations

Laser operation requires independent hardware interlocks, an emergency stop, a protected working area, and target confirmation before emission. Outdoor use also places additional demands on perception: strong sunlight can reduce active RGB-D measurements, while wind and vibration can change branch position during aiming. Repeated visual localization, motion checks, and independent laser-safety control are therefore necessary for reliable field operation.

Several limitations remain. The VOID results show unresolved large-error outliers, and self-supervised view synthesis may degrade under non-rigid branch motion, rolling-shutter artifacts, or platform vibration. Future work will focus on outlier suppression, embedded deployment, and larger outdoor trials.

5. Conclusions

This paper presented RG-RGD, a task-triggered local RGB-D depth-refinement framework for robotic 3D manipulation of small targets. The method combines self-play benefit-driven foveation, residual prediction with Bayesian measurement fusion, UACSPN propagation, and IMU-assisted self-supervised training. On the public VOID benchmark, RG-RGD achieved 24.95 mm MAE and 10.85 iMAE. On self-collected London plane fruit-ball data, BFS-SOFA improved ROI geometric error more strongly than full-image error, while IMU conditioning reduced view-consistency losses relative to vision-only PoseNet. Runtime profiling showed 72.70 ms mean end-to-end Python-pipeline latency and 78.05 ms P95 latency at 320×320 resolution on an RTX 3070 Ti laptop GPU. The qualitative laser-ablation use case illustrates how refined local geometry can support target localization, branch filtering, cutting-point selection, and gimbal-based execution. These results suggest that task-oriented local RGB-D refinement is a practical perception component for robotic manipulation.

Author Contributions: Conceptualization, D.N., J.H. and B.S.; methodology, B.S.; software, B.S.; validation, B.S., Y.G., M.Y., F.Z. and Z.L.; formal analysis, B.S.; investigation, B.S., Y.G., M.Y., F.Z. and Z.L.; resources, D.N. and J.H.; data curation, B.S.; writing—original draft preparation, B.S.; writing—review and editing, D.N., J.H., Y.G., M.Y., F.Z. and Z.L.; visualization, B.S.; supervision, D.N. and J.H.; project administration, D.N.; funding acquisition, D.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 52571377; the National Key Research and Development Program of China, grant number 2023YFC2809804; and the Fundamental Research Funds for the Central Universities, grant numbers 3132023513 and 3132025120.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The public VOID dataset used in this study is available from its official repository. The self-collected RGB-D and IMU data generated during the current study are available from the corresponding author upon reasonable request, subject to institutional approval and reasonable academic use. Source code and reproduction guidelines are available at https://github.com/sakura729/RG-RGD-Depth-Refinement (accessed on 23 July 2026).

Acknowledgments: The authors would like to thank the laboratory members who assisted with prototype construction and data collection.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

BFS-SOFA	Benefit-driven Foveated Scale head with Small-Object Focused Attention
BMF	Bayesian measurement fusion
GBPN	Gaussian belief-propagation network
IMU	Inertial measurement unit
iMAE/iRMSE	Inverse mean absolute error/inverse root mean square error
MAE	Mean absolute error
P95/P99	Ninety-fifth/ninety-ninth percentile; quantity and units are specified in each table
RGB-D	Red–green–blue plus depth
RG-RGD	Residual-Gated RGB-D Depth Refinement
RMSE	Root mean square error
ROI	Region of interest
SOFA	Small-object focused attention
SSIM	Structural similarity index measure
UACSPN	Uncertainty-aware convex spatial propagation network
VOID	Visual Odometry with Inertial and Depth

References

1. Alvarez-Lopez, S.; Fernandez-Gonzalez, M.; Sanchez Espinosa, K.C.; Amigo, R.; Rodriguez-Rajo, F.J. London Plane Tree Pollen and Pla A 1 Allergen Concentrations Assessment in Urban Environments. *Forests* **2022**, *13*, 2089. [CrossRef]
2. Carinanos, P.; Ruiz-Penuela, S.; Valle, A.M.; Diaz de la Guardia, C. Assessing pollination disservices of urban street-trees: The case of London-plane tree (*Platanus x hispanica* Mill. ex Muenchh). *Sci. Total Environ.* **2020**, *737*, 139722. [CrossRef] [PubMed]
3. Navone, A.; Martini, M.; Chiaberge, M. Autonomous robotic pruning in orchards and vineyards: A review. *Smart Agric. Technol.* **2025**, *12*, 101283. [CrossRef]
4. Nouri, N.; Gharakhani, H. A dedicated review of robotic pruners: Current technologies, challenges, and future directions. *Comput. Electron. Agric.* **2025**, *239*, 111013. [CrossRef]
5. You, A.; Parayil, N.; Krishna, J.G.; Bhattarai, U.; Sapkota, R.; Ahmed, D.; Whiting, M.; Karkee, M.; Grimm, C.M.; Davidson, J.R. An autonomous robot for pruning modern, planar fruit trees. *arXiv* **2022**, arXiv:2206.07201.
6. Giang, T.T.H.; Ryoo, Y.-J. Autonomous Robotic System to Prune Sweet Pepper Leaves Using Semantic Segmentation with Deep Learning and Articulated Manipulator. *Biomimetics* **2024**, *9*, 161. [CrossRef] [PubMed]
7. Sarbolandi, H.; Lefloch, D.; Kolb, A. Kinect range sensing: Structured-light versus Time-of-Flight Kinect. *Comput. Vis. Image Underst.* **2015**, *139*, 1–20. [CrossRef]
8. Foix, S.; Alenya, G.; Torras, C. Lock-in Time-of-Flight (ToF) Cameras: A Survey. *IEEE Sens. J.* **2011**, *11*, 1917–1926. [CrossRef]
9. Ma, F.; Karaman, S. Sparse-to-Dense: Depth Prediction from Sparse Depth Samples and a Single Image. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018; pp. 4796–4803. [CrossRef]

10. Cheng, X.; Wang, P.; Yang, R. Depth Estimation via Affinity Learned with Convolutional Spatial Propagation Network. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 108–125. [CrossRef]
11. Tang, J.; Tian, F.-P.; Feng, W.; Li, J.; Tan, P. Learning Guided Convolutional Network for Depth Completion. *IEEE Trans. Image Process.* **2021**, *30*, 1116–1129. [CrossRef] [PubMed]
12. Park, J.; Joo, K.; Hu, Z.; Liu, C.-K.; Kweon, I.S. Non-local Spatial Propagation Network for Depth Completion. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 120–136. [CrossRef]
13. Hu, M.; Wang, S.; Li, B.; Ning, S.; Fan, L.; Gong, X. PENet: Towards Precise and Efficient Image Guided Depth Completion. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 13656–13662. [CrossRef]
14. Li, A.; Yuan, Z.; Ling, Y.; Chi, W.; Zhang, S.; Zhang, C. A Multi-Scale Guided Cascade Hourglass Network for Depth Completion. In Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass Village, CO, USA, 1–5 March 2020; pp. 32–40. [CrossRef]
15. Ryu, K.; Lee, K.-I.; Cho, J.; Yoon, K.-J. Scanline Resolution-Invariant Depth Completion Using a Single Image and Sparse LiDAR Point Cloud. *IEEE Robot. Autom. Lett.* **2021**, *6*, 6961–6968. [CrossRef]
16. Kam, J.; Kim, J.; Kim, S.; Park, J.; Lee, S. CostDCNet: Cost Volume Based Depth Completion for a Single RGB-D Image. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 257–274. [CrossRef]
17. Liu, T.Y.; Agrawal, P.; Chen, A.; Hong, B.-W.; Wong, A. Monitored Distillation for Positive Congruent Depth Completion. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 35–53. [CrossRef]
18. Zhang, Y.; Guo, X.; Poggi, M.; Zhu, Z.; Huang, G.; Mattoccia, S. CompletionFormer: Depth Completion with Convolutions and Vision Transformers. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 18527–18536. [CrossRef]
19. Wang, Y.; Li, B.; Zhang, G.; Liu, Q.; Gao, T.; Dai, Y. LRRU: Long-short Range Recurrent Updating Networks for Depth Completion. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 9388–9398. [CrossRef]
20. Chen, D.; Huang, T.; Song, Z.; Deng, S.; Jia, T. AGG-Net: Attention Guided Gated-convolutional Network for Depth Image Completion. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 8819–8828. [CrossRef]
21. Wong, A.; Cicek, S.; Soatto, S. Learning Topology from Synthetic Data for Unsupervised Depth Completion. *IEEE Robot. Autom. Lett.* **2021**, *6*, 1495–1502. [CrossRef]
22. Wong, A.; Fei, X.; Tsuei, S.; Soatto, S. Unsupervised Depth Completion from Visual Inertial Odometry. *IEEE Robot. Autom. Lett.* **2020**, *5*, 1899–1906. [CrossRef]
23. Zhou, T.; Brown, M.; Snavely, N.; Lowe, D.G. Unsupervised Learning of Depth and Ego-Motion from Video. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6612–6619. [CrossRef]
24. Godard, C.; Mac Aodha, O.; Firman, M.; Brostow, G.J. Digging Into Self-Supervised Monocular Depth Estimation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3827–3837. CrossRef]
25. Watson, J.; Mac Aodha, O.; Prisacariu, V.; Brostow, G.J.; Firman, M. The Temporal Opportunist: Self-Supervised Multi-Frame Monocular Depth. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 1164–1174. [CrossRef]
26. Wu, Y.; Liu, T.Y.; Park, H.; Soatto, S.; Lao, D.; Wong, A. AugUndo: Scaling Up Augmentations for Monocular Depth Completion and Estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024; pp. 274–293. [CrossRef]
27. Almalioglu, Y.; Turan, M.; Saputra, M.R.U.; de Gusmao, P.P.B.; Markham, A.; Trigoni, N. SelfVIO: Self-supervised deep monocular Visual-Inertial Odometry and depth estimation. *Neural Netw.* **2022**, *150*, 119–136. [CrossRef] [PubMed]
28. Wofk, D.; Ranftl, R.; Muller, M.; Koltun, V. Monocular Visual-Inertial Depth Estimation. In Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA), London, UK, 29 May–2 June 2023; pp. 6095–6101. [CrossRef]
29. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial Transformer Networks. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2015; Volume 28.
30. Mnih, V.; Heess, N.; Graves, A.; Kavukcuoglu, K. Recurrent Models of Visual Attention. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2014; Volume 27.

31. Rao, Y.; Zhao, W.; Liu, B.; Lu, J.; Zhou, J.; Hsieh, C.-J. DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 13937–13949.
32. Ryoo, M.S.; Piergiovanni, A.J.; Arnab, A.; Dehghani, M.; Angelova, A. TokenLearner: Adaptive Space-Time Tokenization for Videos. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 12786–12797.
33. Yang, J.; Li, C.; Zhang, P.; Dai, X.; Xiao, B.; Yuan, L.; Gao, J. Focal Attention for Long-Range Interactions in Vision Transformers. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 30008–30022.
34. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 3–7 May 2021.
35. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002. [CrossRef]
36. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
37. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015; pp. 234–241. [CrossRef]
38. Hatamizadeh, A.; Tang, Y.; Nath, V.; Yang, D.; Myronenko, A.; Landman, B.; Roth, H.R.; Xu, D. UNETR: Transformers for 3D Medical Image Segmentation. In Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2022; pp. 1748–1758. [CrossRef]
39. Lenz, C.; Menon, R.; Schreiber, M.; Jacob, M.P.; Behnke, S.; Bennewitz, M. HortiBot: An Adaptive Multi-Arm System for Robotic Horticulture of Sweet Peppers. In Proceedings of the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Abu Dhabi, United Arab Emirates, 14–18 October 2024; pp. 31–38. [CrossRef]
40. Giang, T.T.H.; Ryoo, Y.-J. Pruning Points Detection of Sweet Pepper Plants Using 3D Point Clouds and Semantic Segmentation Neural Network. *Sensors* **2023**, *23*, 4040. [CrossRef] [PubMed]

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