Manifold-Guided Stereo-Monocular Refinement for Endoscopic Stereo Disparity Estimation

Jiewen Yang[1] **Yi Zhong**[2] **Bin Pu**[3*] **Xingbo Dong**[2] **Ming Zheng**[5]
Qika Lin[4] **Kai He**[4] **Xuechen Zhang**[1*]

[1]The Hong Kong University of Science and Technology [2]Anhui University
[3]Hunan University [4]National University of Singapore [5]Anhui Normal University

jyangcu@connect.ust.hk wa2314072@stu.ahu.edu.cn pubin@hnu.edu.cn
xingbo.dong@ahu.edu.cn mzheng@ahnu.edu.cn qikalin@foxmail.com
kai_he@nus.edu.sg xhangem@connect.ust.hk

Abstract

Dense stereo disparity estimation provides the geometric basis for 3D perception in computer-assisted endoscopy. Endoscopic stereo matching is difficult because reliable tissue correspondences often appear next to weak-texture, specular, occluded, or deformed regions within the same frame. Existing stereo and stereo-monocular methods improve cost aggregation or introduce monocular priors, but they often treat monocular guidance as a global feature source or an independently predicted candidate. Without evaluating the local support of the current stereo match during refinement, the update can reinforce corrupted correspondences in ambiguous regions or allow monocular priors to override reliable metric matches. To tackle these problems, we introduce Manifold-Stereo, a reliability-gated latent-flow framework for endoscopic stereo disparity estimation. The method refines a monocular disparity candidate through a conditional latent residual flow on an image-aligned correction space induced by stereo cost-volume geometry, then calibrates the corrected candidate to the current stereo disparity scale. A geometric consistency gate computed from left-right feature alignment regulates the candidate contribution during recurrent refinement, preserving stereo estimates in well-matched regions while supplying monocular context where local matching is ambiguous. Experiments on in-domain SCARED and zero-shot SERV-CT and Hamlyn show strong average disparity accuracy and competitive thresholded outlier rates against recent stereo and endoscopic baselines. Code is available at: https://github.com/Jiewen-Yang/Manifold-Stereo.

1 Introduction

Accurate dense stereo disparity estimation is central to 3D perception in minimally invasive endoscopy [1–3]. Reliable disparity supports anatomical surface reconstruction, surgical navigation, augmented reality guidance, and robot-assisted manipulation [1, 4]. Unlike natural-image stereo, endoscopic stereo is captured at close range inside illuminated surgical cavities, where deforming wet tissue, fluid, specular reflection, smoke, and instruments make local correspondence unstable [5, 6].

The key challenge is not a uniform domain shift but spatially uneven correspondence reliability within each frame [2, 3, 7]. Textured tissue and stable boundaries can provide strong stereo evidence, whereas weak texture, highlights, motion blur, occlusions, and contact regions can corrupt neighboring matches [5–7]. This endoscopic structure calls for a refinement model that preserves well-supported

*Corresponding Author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

metric correspondence while using anatomical context [8–10] only where local stereo evidence is ambiguous.

Existing endoscopic geometry methods address unreliable correspondence through complementary mechanisms. Classical stereo pipelines [1, 11, 12] define matching costs and regularizers. Learning-based methods [2, 7, 13–18] learn correspondence representations, context aggregation, or self-supervised reconstruction objectives. Recent approaches further introduce monocular priors [15–19] to provide anatomical context when stereo evidence is weak. However, many refinement pipelines use monocular guidance as a global feature source or an independently generated candidate [15–19], without testing whether the current stereo estimate is locally supported by left-right alignment. This leaves two failure modes. Stereo-only updates may reinforce corrupted matches in weak-texture, specular, or occluded regions, while monocular priors may overwrite reliable calibrated matches. A suitable refinement model should preserve calibrated stereo where left-right alignment is reliable, introduce monocular context only where stereo support is weak, and express monocular corrections in an image-aligned representation anchored to stereo evidence.

Motivated by this endoscopic reliability structure, we introduce Manifold-Stereo, a reliability-gated latent-flow framework for endoscopic stereo disparity estimation. As shown in Figure 1, the method treats stereo and monocular estimates as disparity candidates whose roles are determined by local stereo support. Beyond transferring monocular features into descriptor and cost aggregation, Manifold-Stereo explicitly treats the monocular prediction as a disparity candidate whose correction and contribution are controlled by stereo geometry. It first corrects the monocular candidate under stereo cost-volume geometry, then uses local left-right consistency to regulate the corrected candidate during recurrent stereo refinement.

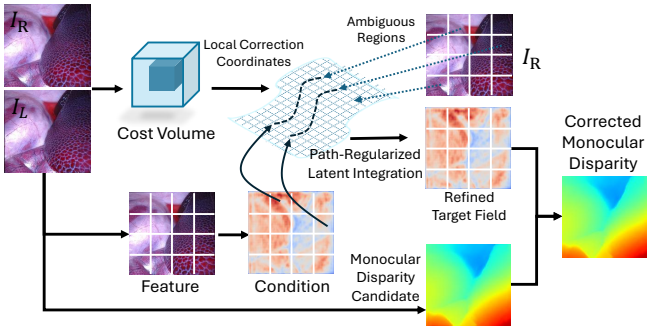

Figure 1: Motivation for reliability-aware stereo-monocular refinement. Stereo matching provides metric correspondence evidence where left-right alignment is reliable, while a locally refined monocular candidate can support regions with ambiguous correspondence.

The first step makes monocular information compatible with calibrated stereo. Monocular priors can be informative on homogeneous tissue or specular regions, but unconstrained monocular correction can conflict with precise stereo matches. Manifold-Stereo therefore represents monocular correction as a cost-volume-conditioned latent trajectory. The cost-volume representation defines an image-aligned correction space around the current correspondence evidence, and the latent vector field produces a regularized residual disparity update within this stereo-conditioned neighborhood. The result is a corrected monocular candidate tied to the geometry observed by the stereo pair.

The second step determines the spatial contribution of this corrected candidate. Because stereo support varies from textured tissue to specular, occluded, or weak-texture regions, Manifold-Stereo estimates local support through left-right feature consistency under the current stereo disparity. The resulting geometric consistency gate regulates the corrected candidate's contribution during recurrent refinement. This design preserves calibrated stereo for well-supported matches and provides controlled monocular guidance where stereo-only refinement would propagate corrupted matches. The main contributions are listed below.

- A reliability-gated stereo-monocular disparity refinement formulation that uses local left-right feature consistency to regulate monocular candidate contribution during recurrent stereo updates.
- A conditional latent residual flow for monocular candidate refinement in a cost-volume-induced local correction space, enabling regularized, location-specific disparity updates before reliability-gated stereo refinement.
- State-of-the-art average disparity accuracy on in-domain SCARED and zero-shot SERV-CT and Hamlyn evaluation, achieving the best aggregate outlier rates on SCARED and SERV-CT and competitive outlier rates on Hamlyn.

2 Related Work

2.1 Stereo Disparity, Monocular Priors, and Foundation Models

Stereo matching has evolved from hand-crafted costs and optimization to deep cost-volume learning. GC-Net [20], PSMNet [21], and GwcNet [22] introduced 3D cost regularization, multi-scale aggregation, and group-wise correlation, respectively. Iterative methods such as RAFT-Stereo [23] refine disparity through recurrent updates [cf. recurrent video modeling in 24], while efficient or generalist stereo models such as AANet [25], HITNet [26], FoundationStereo [27], and Monster [28] improve efficiency, cross-domain robustness, or stereo-monocular generalization. These models strengthen correspondence estimation, but their update rules usually treat the cost volume as a sufficient reliability source and do not explicitly decide when an external monocular candidate should be trusted over the current stereo state.

Monocular foundation depth models such as Depth Anything methods [29, 30] provide semantic and anatomical priors for ambiguous regions. However, monocular predictions are scale ambiguous and can be geometrically inconsistent with a calibrated stereo pair. Manifold-Stereo therefore does not use foundation-depth information as an independently trusted prediction or only as a global feature source. It uses the monocular branch to transfer contextual features for stereo matching and to construct an uncalibrated disparity candidate, then refines this candidate under stereo cost-volume geometry and regulates its contribution using stereo reliability. The cost volume induces a local correction space for residual candidate refinement, while a left-right feature-consistency gate controls the corrected candidate's contribution to recurrent stereo updates.

2.2 Endoscopic Stereo Disparity and Geometry-Guided Refinement

Endoscopic stereo disparity estimation differs from natural-scene stereo because tissue often exhibits specular appearance, weak texture, non-Lambertian reflectance, and dynamic deformation. SCARED, Hamlyn, and SERV-CT [3, 6, 31] enable evaluation under these conditions, either directly in disparity space or through depth converted from calibrated disparity. Prior work addresses reflection artifacts through detection or inpainting [5, 7], uses stereo-pair geometry, anatomical priors, or monocular priors for tissue with weak texture [4, 15–17], and exploits temporal cues in video endoscopy [7, 18]. These methods improve robustness, but many do not use left-right feature consistency to regulate the effect of monocular refinement candidates on recurrent stereo disparity updates.

Our refinement formulation is also related to latent flows and local geometric parameterizations [32]. Continuous vector fields and neural ODEs parameterize transformations as integrated trajectories [33], while flow matching studies supervised vector fields for distribution transport [34]. Manifold and representation learning motivate local coordinate systems and decoder-induced geometries [35–37]. In Manifold-Stereo, the latent flow is not a distributional flow-matching objective. It is a regularized residual trajectory over image-aligned cost-volume features. The trajectory corrects a monocular candidate before geometric consistency gating regulates the candidate's contribution to recurrent stereo refinement.

3 Methodology

We present Manifold-Stereo, a stereo disparity estimation framework for endoscopic scenes. Stereo correspondence reliability is spatially uneven. Calibrated matching constrains textured tissue, whereas occluded, specular, or weak-texture regions require contextual support. Given a rectified stereo pair (I_L, I_R), the model predicts a dense disparity map D. Manifold-Stereo refines monocular candidates within a cost-volume-induced local correction space and incorporates them into recurrent stereo updates through a geometric consistency gate. Here, calibrated stereo refers to the known camera geometry of the rectified pair; stereo evidence denotes matching information, and stereo support denotes agreement under the current disparity.

3.1 Reliability-Aware Refinement Formulation

Figure 2 summarizes the pipeline. Manifold-Stereo uses calibrated stereo matching and monocular context as spatially adaptive evidence sources rather than assigning fixed global weights. This distinction is essential in endoscopy, where reliable and unreliable correspondences often coexist within

a single frame. Textured tissue can provide strong metric stereo evidence, whereas nearby specular or weak-texture regions may require contextual support from a monocular estimate. The refinement problem is formulated as cost-volume-guided correction with reliability-gated candidate contribution. A stereo-conditioned latent trajectory corrects the monocular candidate, and local geometric reliability determines where the corrected candidate contributes to recurrent stereo refinement.

The formulation contains two coupled components. First, latent-flow refinement corrects the monocular disparity candidate in a spatial latent field conditioned on stereo cost-volume geometry. The cost-volume encoder and residual decoder define an operational local correction space, where image-aligned latent coordinates are extracted from cost-volume geometry and latent displacements decode to disparity residuals. We use the term "flow" for the parameter-shared time-one map of this learned residual vector field, not for distributional flow matching. Endpoint anchoring and path-energy regularization discourage departures from this cost-volume-induced neighborhood. Second, reliability gating controls the corrected candidate's contribution according to the local support of the current stereo estimate. The gate is computed from left-right feature alignment under the current stereo disparity, providing a local reliability score for metric correspondence. Candidate construction supplies the evidence required by this reliability-gated latent-flow refinement.

3.2 Stereo and Monocular Candidate Construction

The first stage constructs the evidence used by cost-volume-guided refinement. The stereo branch extracts left and right matching descriptors and builds a calibrated cost volume, while the monocular branch provides anatomical context and an uncalibrated disparity-like candidate. Monocular features guide descriptor and cost aggregation, but the stereo cost volume remains the source of metric correspondence evidence and the anchor for the local correction space. Layer-level details are given in Appendix VII.

From the combined $\frac{1}{4}$-resolution left and right matching descriptors $\tilde{f}_L$ and $\tilde{f}_R$, we build a group-wise correlation volume as

$$C_g(x, y, d) = \frac{1}{|\mathcal{C}_g|} \sum\nolimits_{c \in \mathcal{C}_g} \tilde{f}_L^c(x, y) \tilde{f}_R^c(x - d, y), \tag{1}$$

where $\mathcal{C}_g$ denotes the channel set of group g, and d indexes disparity candidates. Group-wise correlation retains multiple subspace-specific matching responses.

A monocular-attended 3D aggregation network $\mathcal{A}_{\text{cost}}$ regularizes the cost volume, producing a geometric encoding volume V_{geo}. A classifier and soft argmax produce the initial stereo disparity D_{init}, while the frozen monocular branch predicts an uncalibrated disparity-like candidate D_{mono}. We do not assume this candidate has metric scale or offset; it is aligned to the current stereo estimate before reliability-gated recurrent refinement.

For latent candidate refinement, V_{geo} is encoded into a spatial latent field $Z_0 \in \mathbb{R}^{B \times L \times \frac{H}{4} \times \frac{W}{4}}$. Pooling is restricted to the disparity dimension, preserving image alignment and location-specific coordinates for the latent residual-flow update.

3.3 Conditional Latent Residual Flow in a Local Correction Space

The latent residual flow refines the monocular disparity candidate before it contributes to the recurrent stereo state. It generates a spatially varying residual correction relative to that candidate using the local geometry encoded by the stereo cost volume. The geometric cost volume is mapped to a spatial latent field by a $3 \times 3 \times 3$ convolution followed by pooling along the disparity dimension only. Each image location p therefore has a latent coordinate $Z_0(p) \in \mathbb{R}^L$ aligned with local stereo cost-volume evidence.

Let ψ_θ denote the residual decoder that maps a spatial latent field to a disparity offset. Around the cost-volume latent state Z_0, we define the local correction space for the monocular candidate as

$$\mathcal{C}_\theta(D_{\text{mono}}, V_{\text{geo}}) = \{D_{\text{mono}} + \psi_\theta(Z) : Z \in \mathcal{U}(Z_0)\}, \tag{2}$$

where $\mathcal{U}(Z_0)$ is a bounded neighborhood of the cost-volume-induced latent field. *Manifold* refers to this decoder-induced local correction geometry, without asserting a global anatomical manifold or constant-rank guarantee. Although Z_0 contains transferred monocular features, it encodes stereo

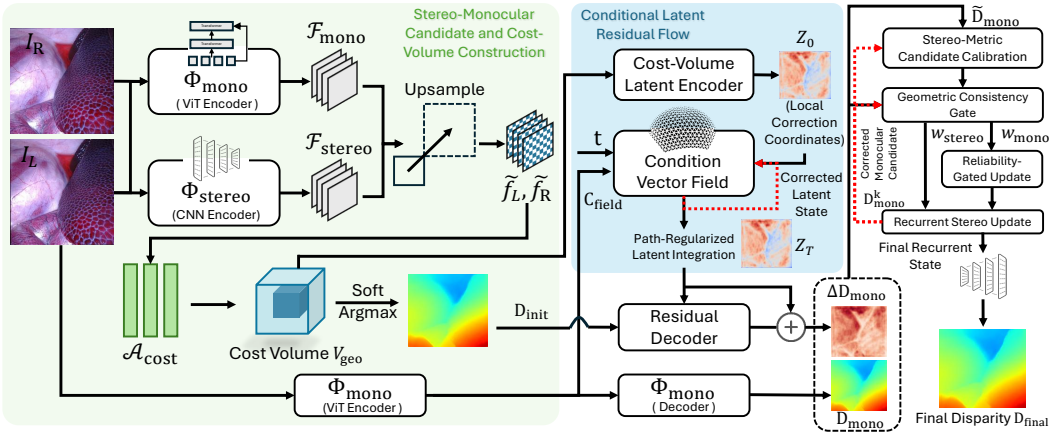

Figure 2: **Pipeline of Manifold-Stereo for endoscopic stereo disparity estimation.** Given a rectified stereo pair, stereo and monocular disparity candidates are constructed from cost-volume geometry and monocular context. A conditional latent residual flow refines the monocular candidate in an image-aligned local correction space, and geometric consistency-gated recurrent refinement controls its contribution to calibrated stereo updates.

matching geometry; D_{mono} supplies the explicit base estimate. The decoder ψ_θ maps the refined latent field to a $\frac{1}{4}$-resolution residual, with their sum supervised by disparity losses. Appendix IX details these complementary roles.

Let $\bar{Z} = \text{sg}(Z_0)$ denote the stop-gradient reference state (latent anchor). This reference centers the local correction neighborhood, while disparity supervision selects useful decoded corrections. A conditioning field C_{field} is extracted from the right matching descriptor by a 1×1 convolution and defines a conditional vector field $v_\theta(t, Z_t, C_{\text{field}})$ at each image location.

At inference time, the refined latent state is obtained by integrating the learned vector field over $t \in [0, T]$ with $T = 1$,

$$Z_T = Z_0 + \int_0^T v_\theta(t, Z_t, C_{\text{field}})\, dt. \tag{3}$$

Integration is performed over flattened spatial latent coordinates, with the latent tensor reshaped back to the image lattice after each step. In implementation, we use an 8-step RK4 solver on the $\frac{1}{4}$-resolution latent field. This should be read as an ODE-style, parameter-shared latent residual refiner: the same vector field is evaluated at multiple pseudo-time locations, and the solver itself introduces no additional learnable parameters beyond the vector field and residual decoder. The one-step ablation in Table 6 keeps the same latent encoder, conditioning field, vector field, and decoder, but removes this multi-step trajectory. The refined latent state is decoded to a disparity offset $\Delta D_{\text{mono}} = \psi_\theta(Z_T)$, yielding $\tilde{D}_{\text{mono}} = D_{\text{mono}} + \Delta D_{\text{mono}}$. Appendix V formalizes the decoder-induced local correction space, path-energy regularization, and reliability-weighted contribution control.

3.4 Geometric Consistency-Gated Recurrent Refinement

The geometric consistency gate regulates the corrected monocular candidate in the reliability-gated latent-flow formulation. It determines the candidate's spatial contribution to recurrent stereo refinement. The recurrent stereo state is initialized by D_{init} and updated from sampled cost-volume evidence. Early iterations use stereo-only updates to establish a calibrated metric estimate. Later iterations introduce the monocular candidate aligned to the stereo scale, with contribution weights determined by local left-right feature consistency rather than fixed, location-independent stereo-monocular mixing.

The corrected monocular representation is weighted by local geometric consistency. Let D^k_{stereo} denote the current stereo-branch disparity and D^k_{mono} the monocular-branch disparity aligned to the stereo scale and used in late recurrent iterations. Given left and right features f_L and f_R, we first warp the right feature to the left view using D^k_{stereo} and compute the feature-alignment residual. The

geometric consistency gate is then defined as

$$\mathcal{G}_{\text{stereo}} = \exp\left(-\frac{\|f_L - \text{Warp}(f_R, D^k_{\text{stereo}})\|^2}{C\tau}\right), \tag{4}$$

where C is the number of feature channels and $\tau = 0.1$. This score measures left-right feature alignment under the current stereo disparity. We set $w_{\text{stereo}} = \mathcal{G}_{\text{stereo}}$ and $w_{\text{mono}} = 1 - \mathcal{G}_{\text{stereo}}$. Let V_{stereo} and V_{mono} denote the stereo and monocular update inputs produced from D^k_{stereo} and D^k_{mono}, such as sampled geometric features and feature-alignment residuals. The gated update input is $V_{\text{fused}} = w_{\text{stereo}} V_{\text{stereo}} + w_{\text{mono}} V_{\text{mono}}$. Thus, $\mathcal{G}_{\text{stereo}}$ acts as an empirical confidence score for local left-right feature support. Appendix V interprets this rule as the closed-form solution of a local reliability-weighted least-squares contribution problem.

Since the monocular candidate has an unknown relative scale and offset, *per-pair stereo-scale alignment* is applied before late stereo-monocular refinement. The affine transform $s\tilde{D}_{\text{mono}} + b$ is estimated by least-squares fitting to the current stereo disparity D^k_{stereo}, with a small Tikhonov stabilizer. This operation adjusts the candidate, not camera parameters, and uses no ground truth. The fitting mask excludes non-positive stereo values and unstable monocular values; Appendix IV gives the criteria. Later iterations maintain separate stereo and monocular states, each receiving its gated feature and the other branch's gated feature as context. The complementary weight controls contribution; it does not assert monocular correctness when stereo is unreliable. Appendix XI explains the gate and its diagnostics.

The final full-resolution disparity is recovered after recurrent refinement with an output residual decoder. Appendix VII provides details.

3.5 Training Objective

The model is trained end-to-end with $\mathcal{L} = \mathcal{L}_{\text{seq}} + \lambda_2 \mathcal{L}_{\text{flow}} + \lambda_3 \mathcal{L}_{\text{geo}}$, where $\lambda_2 = 10^{-3}$ and $\lambda_3 = 10^{-1}$. The shared recurrent sequence loss $\mathcal{L}_{\text{seq}}$ supervises the initial cost-volume disparity and all recurrent predictions. Late-stage monocular-branch predictions are inserted into the same recurrent sequence with the same weighting schedule. Appendix IV gives details. The latent regularizer is

$$\mathcal{L}_{\text{flow}} = \|Z_T - \bar{Z}\|_2^2 + \alpha \sum\nolimits_{n=0}^{N-1} \|Z_{n+1} - Z_n\|_2^2, \tag{5}$$

where $\{Z_n\}_{n=0}^N$ denotes the discrete integration states with $Z_N = Z_T$. The first term anchors the endpoint to the local cost-volume embedding, and the second penalizes unstable trajectories in the local correction space, with $\alpha = 0.5$. Decoded disparity losses determine the refinement direction. This loss corresponds to the endpoint consistency and path-energy surrogate discussed in Appendix V. The geometry-aware consistency term uses the feature-consistency score from the gate to weight auxiliary agreement between the final prediction and the detached monocular-branch geometric candidate,

$$\mathcal{L}_{\text{geo}} = \frac{\sum_{p \in \Omega_v} w(p)\ \text{smooth}_{\ell_1}\left(D_{\text{final}}(p) - D_{\text{geo}}(p)\right)}{\sum_{p \in \Omega_v} w(p) + \epsilon}, \tag{6}$$

where $D_{\text{geo}} = \text{sg}(D^{\text{up}}_{\text{mono,ref}})$ and $w(p)$ increases with feature uncertainty and decreases with stereo consistency. The weighting rule is given in Appendix IV.

4 Experiment

4.1 Datasets

SCARED [3] contains rectified stereo images of porcine cadavers captured with the da Vinci Xi surgical system and structured-light 3D ground truth at 1280×1024 resolution. The dataset includes nine subsets, D1–D9. We follow the training, validation, and testing split used in prior work [7]. **SERV-CT [31]** contains 16 porcine torso stereo sets captured by an O-arm™ system at 720×576 resolution. **Hamlyn [6]** contains surgical videos with partial geometry annotations, tissue deformation, tool interaction, and smoke occlusion. Following [2, 6], we use 22 Hamlyn videos for zero-shot evaluation. The predicted left-view disparity is the primary output. For fair comparison, all baselines are reproduced under the same dataset splits, training-validation protocol, and metrics as our method.

Table 1: Evaluation of disparity and depth estimation on SCARED across all test subsets.

Methods	Test subset 1 (Depth MAE: mm ↓)					Test subset 2 (Depth MAE: mm ↓)					mean Depth MAE (mm)↓	D1 (%)↓	Bad1 (%)↓	Bad2 (%)↓	Bad3 (%)↓	EPE (px)↓
	K0	K1	K2	K3	K4	K0	K1	K2	K3	K4						
GwcNet [22]	8.60	2.63	1.49	2.06	0.66	4.25	1.00	3.41	1.44	0.35	2.59	27.36	74.63	52.57	35.66	3.91
ACVNet [38]	8.56	2.59	1.49	2.08	0.68	4.23	0.96	3.46	1.44	**0.33**	2.58	24.91	72.92	53.51	36.50	3.95
RAFT-Stereo [23]	8.45	2.51	1.75	2.03	0.73	**4.14**	**0.89**	3.06	1.52	0.68	2.58	28.46	74.28	53.58	38.80	4.14
IGEV-Stereo [39]	8.27	2.35	1.73	2.18	**0.56**	4.41	0.97	3.22	1.54	0.37	2.56	25.67	71.56	50.29	35.52	3.88
Selective-Stereo [40]	8.44	2.44	1.62	2.27	0.70	4.30	0.94	3.34	1.78	0.66	2.65	27.64	74.81	51.87	36.92	4.12
MSDESIS [41]	8.42	2.59	2.05	3.02	1.01	4.68	1.18	3.34	1.61	0.46	2.84	29.17	77.39	53.34	38.81	4.48
Shi *et al*. [42]	7.61	**2.10**	1.97	2.66	0.65	4.75	1.18	2.96	1.71	0.36	2.60	26.45	73.88	52.98	37.16	4.01
DynamicStereo [43]	8.64	2.50	1.66	**1.95**	-	4.23	1.03	3.76	1.66	-	3.18	31.82	79.46	57.21	42.16	4.38
Monster [28]	8.10	2.41	1.63	2.59	1.04	4.37	1.08	2.93	1.67	0.51	2.68	30.39	77.54	53.51	37.27	3.51
RRESM [44]	7.62	2.31	1.66	2.28	1.02	4.42	1.16	2.80	2.12	0.47	2.59	26.78	73.42	51.65	36.93	4.02
StereoMamba [45]	8.57	2.61	**1.46**	2.03	0.77	4.19	0.92	3.07	1.46	0.41	2.55	26.35	75.22	53.09	36.05	3.39
Ours	**7.35**	2.20	1.85	2.71	0.98	**4.14**	0.99	**2.69**	**1.43**	0.91	**2.53**	**24.33**	**71.52**	**49.68**	**35.03**	**3.05**

*$\mathbf{K}n$ represents the n-th keyframe of each subset. † DynamicStereo [43] is specifically designed for sequential inputs and cannot be applied in $K4$ as this subset contains only single-frame stereo pairs.

Table 2: Zero-shot SERV-CT results. DynamicStereo requires sequences and is inapplicable to these static pairs.

Model	EPE(px)↓	D1(%)↓	Bad1(%)↓	Bad2(%)↓	Bad3(%)↓
GwcNet [22]	5.81	33.92	79.37	56.32	40.93
ACVNet [38]	6.36	29.15	70.88	49.80	33.87
RAFT-Stereo [23]	3.24	19.70	67.20	41.85	27.12
IGEV [39]	6.58	36.91	77.28	57.53	41.76
Selective-Stereo [40]	5.21	34.22	79.62	47.90	35.97
MSDESIS [41]	5.74	28.26	78.81	55.44	37.79
Shi *et al*. [42]	7.07	41.35	81.09	63.17	49.22
Monster [28]	4.68	26.48	70.06	44.71	37.93
RRESM [44]	3.11	19.60	78.22	51.73	30.99
StereoMamba [45]	2.63	14.86	63.38	37.82	22.94
Ours	**1.70**	**8.96**	**52.89**	**24.67**	**13.22**

Table 3: Zero-shot Hamlyn results. All methods use SCARED-selected checkpoints without target-data fine-tuning.

Model	EPE(px)↓	D1(%)↓	Bad1(%)↓	Bad2(%)↓	Bad3(%)↓
GwcNet [22]	8.31	38.67	75.37	63.86	51.26
ACVNet [38]	8.85	42.59	66.88	57.34	44.20
RAFT-Stereo [23]	4.71	30.54	57.59	39.48	31.56
IGEV [39]	7.39	42.66	64.28	58.07	46.09
Selective-Stereo [40]	6.60	38.11	61.68	52.87	41.34
MSDESIS [41]	8.24	36.01	74.81	62.98	48.12
Shi *et al*. [42]	9.56	49.10	77.09	70.71	59.55
DynamicStereo [43]	8.45	35.87	71.20	59.64	46.83
Monster [28]	7.57	37.22	73.68	56.15	45.99
RRESM [44]	4.44	**26.91**	**52.90**	**35.04**	28.02
StereoMamba [45]	4.98	28.55	55.12	36.91	29.57
Ours	**4.24**	27.49	57.19	35.13	**27.96**

4.2 Training and Evaluation Setup

Manifold-Stereo constructs stereo and monocular candidates from a calibrated stereo branch and a ViT-based monocular branch. We train the model end-to-end for 50,000 steps with batch size 8 using AdamW, an initial learning rate of 1×10^{-4}, weight decay 1×10^{-5}, and OneCycleLR over $[5 \times 10^{-5}, 1 \times 10^{-4}]$. Images are resized to 512×256, the disparity range is $[0.5, 192)$, and training uses color jittering with saturation in $[0.7, 1.3]$ and spatial scaling in $[-0.2, 0.5]$. Unless otherwise stated, the cost-volume-guided refinement uses geometric consistency-gated candidate contribution and conditional latent residual flow ($L = 128$), integrated with 8 RK4 steps. Training is performed on two RTX 3090 GPUs. The model selected on the SCARED validation set is used for all reported tests, including zero-shot evaluation on SERV-CT and Hamlyn.

4.3 Evaluation Metrics

We report calibrated depth and disparity metrics. In all tables, MAE (mm) denotes calibrated depth mean absolute error after converting predicted and ground-truth disparities to metric depth; it is not used as a disparity metric. EPE (px) measures average disparity error in pixels, while Bad-t reports the percentage of valid pixels with absolute disparity error above t pixels. D1 follows the stereo benchmark convention and counts large outliers under the dataset-specific threshold. For SCARED, we report the established per-keyframe and mean Depth MAE in millimeters under the calibrated evaluation protocol. Appendix III defines all metrics.

4.4 Quantitative Analysis

In-domain evaluation on SCARED. Table 1 reports the in-domain SCARED comparison under the same split protocol. Manifold-Stereo obtains the best mean Depth MAE (2.53 mm), D1 (24.33%), Bad1 (71.52%), Bad2 (49.68%), Bad3 (35.03%), and EPE (3.05 px). These are aggregate gains; individual keyframes are not uniformly best. Compared with the strongest prior entries, it improves the EPE of StereoMamba from 3.39 px to 3.05 px, reduces D1 from ACVNet's 24.91% to 24.33%, and improves the thresholded outlier rates over the best competing Bad1, Bad2, and Bad3 scores. This metric pattern is consistent with the design goal of Manifold-Stereo. The best EPE indicates

lower average disparity error, the best D1 and Bad-t scores indicate fewer outlier pixels across both tight and larger thresholds, and the best mean Depth MAE shows that these gains remain beneficial under calibrated metric evaluation. These results suggest that reliability-gated candidate contribution can suppress ambiguous-region correspondence failures while preserving the locally calibrated stereo matches needed for fine-grained threshold accuracy.

Zero-shot performance on SERV-CT and Hamlyn. Tables 2 and 3 evaluate zero-shot transfer without target-dataset fine-tuning. On SERV-CT, Manifold-Stereo achieves the best performance across all reported metrics, with EPE 1.70 px, D1 8.96%, Bad1 52.89%, Bad2 24.67%, and Bad3 13.22%. On Hamlyn, it obtains the best EPE (4.236 px) and Bad3 (27.96%), and ranks second on D1 (27.49%) and Bad2 (35.13%). RRESM remains stronger on D1, Bad1, and Bad2, while StereoMamba remains stronger on Bad1. This shows strong transfer in average error and larger outlier suppression, although the tighter thresholded metrics remain competitive rather than uniformly best under the stronger video-domain variation in Hamlyn.

This difference follows from the metrics. Bad-t binarizes each pixel at a fixed threshold, whereas EPE and MAE retain error magnitude. Reliability-gated candidate contribution mainly suppresses corrupted matches in weak-texture, specular, or occluded regions, so it can reduce large local errors even when some small-threshold counts are not the best. The zero-shot results therefore support the intended behavior. Manifold-Stereo keeps calibrated stereo as the metric backbone and introduces the corrected monocular candidate selectively where local stereo support is weak.

4.5 Qualitative Analysis

Figures 3 and 4 compare predicted disparity maps under in-domain and zero-shot settings. On SCARED, the compared stereo methods generally recover the organ-scale disparity gradient, but several predictions either smooth local tissue relief or introduce texture-like artifacts around dark borders, specular areas, folded tissue, and instrument-contact regions. Manifold-Stereo produces more coherent continuous surfaces while preserving clearer disparity transitions across folds and boundaries. The visual improvement appears mainly as reduced mottled artifacts and better surface continuity, which matches the quantitative trend of lower average error and improved thresholded outlier rates.

Under the stronger SERV-CT and Hamlyn domain shifts, Manifold-Stereo preserves organ-scale layout while reducing flat transitions and border or weak-texture artifacts observed in several baselines. The resulting surfaces are smoother on homogeneous tissue yet retain clearer low-to-high disparity separation, matching the SERV-CT gains and the Hamlyn improvements in EPE and Bad3. This behavior is consistent with reliability-gated candidate contribution: calibrated stereo remains dominant where correspondence is stable, while the corrected monocular candidate supports locally ambiguous regions.

Figure 5 further shows that latent residual-flow refinement is spatially selective: correction fields concentrate near ambiguous cost-volume responses, tissue boundaries, highlights, and weak-texture regions instead of spreading uniformly. The corrected monocular candidate remains smooth on continuous tissue while preserving disparity transitions, supporting the intended design of cost-volume-guided candidate correction before reliability-gated recurrent refinement. Appendix XI complements these full-frame views with the gate diagnostic and challenging-region analysis.

4.6 Ablation Study

All ablations are conducted on SCARED under the same training and evaluation protocol as the main comparison. The main tables report MAE and EPE for compactness. Appendix VIII provides additional ablations.

Core refinement factors. Table 4 evaluates the complete monocular pathway and its refinement components. The full model obtains MAE 2.53 mm and EPE 3.05 px. *Stereo-only (No Monocular Branch)*, previously abbreviated as *No Monocular Feat.*, removes both monocular features and the disparity candidate; candidate correction, alignment, and gated monocular contribution are bypassed. Its errors are 3.11 mm and 3.90 px. This row measures the whole pathway's contribution, rather than isolating D_{mono} alone. With the candidate retained, removing the local correction mechanism

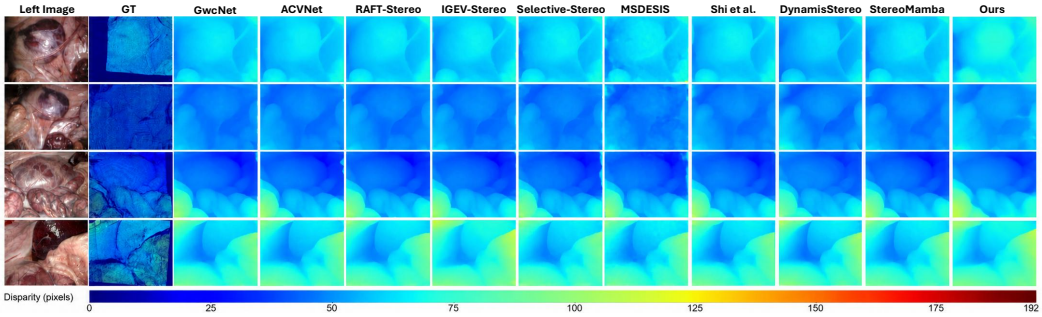

Figure 3: Visualization of disparity estimation on the SCARED dataset.

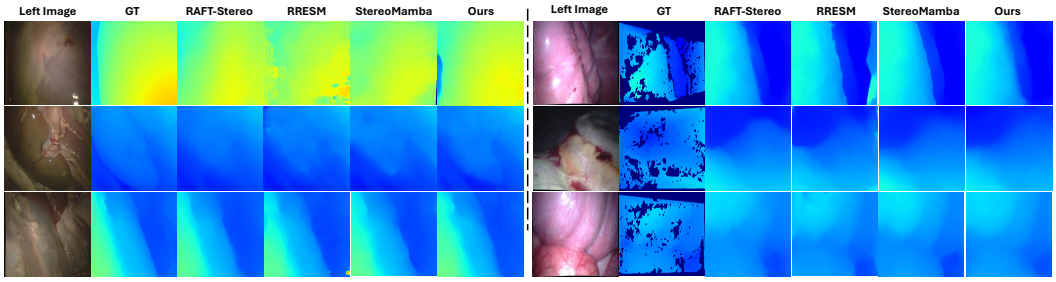

Figure 4: Visualization of zero-shot disparity estimation on SERV-CT (left) and Hamlyn (right).

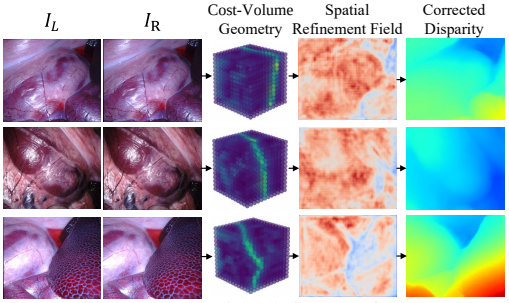

Figure 5: Latent residual-flow refinement produces local, structured corrections that smooth locally ambiguous tissue regions while preserving sharp disparity transitions.

Table 4: Core ablation. N/A denotes operations bypassed without the monocular branch.

Setting	Enabled Factors			MAE (mm)↓	EPE (px)↓
	Monocular Branch	Latent Flow	Geometric Gate		
Full	✓	✓	✓	**2.53**	**3.05**
Stereo-only (No Monocular Branch)	–	N/A	N/A	3.11	3.90
No Correction Space	✓	–	✓	2.60	3.27
No Geometric Gate	✓	✓	–	2.77	3.45

Table 5: Ablation of gated contribution and per-pair stereo-scale alignment.

Contribution/calibration variant	MAE(mm)↓	EPE(px)↓
Full	**2.53**	**3.05**
Uniform Mixing	2.63	3.30
No Stereo-scale Alignment	3.47	4.32

increases the errors to 2.60 mm and 3.27 px, and removing the geometric gate gives 2.77 mm and 3.45 px. Appendix X specifies the component switches and controlled comparisons.

Latent residual refinement. Table 6 evaluates the latent residual-flow component. The one-step variant keeps the same cost-volume latent encoder, conditioning field, vector field, and residual decoder, but evaluates the vector field only once instead of integrating the multi-step trajectory. This change increases the errors from 2.53 mm and 3.05 px to 2.82 mm and 3.55 px, indicating that the gain is not only due to adding a residual prediction head. Removing path-energy regularization further increases the errors to 3.09 mm and 3.84 px, which supports the role of trajectory constraints in keeping monocular corrections within the stereo-conditioned local neighborhood. These results support the latent-flow formulation as a regularized, parameter-shared latent correction process before reliability-gated stereo refinement.

Table 6: Ablation of latent residual refinement on SCARED [3] test set.

Latent refinement variant	MAE(mm)↓	EPE(px)↓
Full	**2.53**	**3.05**
One-step update	2.82	3.55
No path energy	3.09	3.84

Gated candidate contribution and stereo-scale alignment. Table 5 evaluates corrected-candidate contribution during recurrent stereo refinement. Replacing geometric-consistency weighting with uniform mixing increases the errors from 2.53 mm and 3.05 px to 2.63 mm and 3.30 px, indicating that fixed location-independent weights cannot capture spatially varying stereo reliability. Removing per-pair stereo-scale alignment increases the errors to 3.47 mm and 4.32 px, supporting alignment of the candidate before recurrent refinement. Alignment makes the corrected monocular candidate metrically comparable to stereo, while the gate controls its spatial contribution.

5 Conclusion

We presented Manifold-Stereo, an endoscopic stereo disparity framework for spatially varying stereo correspondence reliability. The method combines cost-volume-guided correction with reliability-gated candidate contribution. A conditional latent residual flow refines the monocular disparity candidate in an image-aligned correction space induced by stereo cost-volume geometry, and a geometric consistency gate regulates the corrected candidate during recurrent stereo updates. This formulation preserves metric stereo information in well-matched regions and selectively introduces monocular context in ambiguous regions.

Experiments on SCARED, with zero-shot transfer to SERV-CT and Hamlyn, demonstrate strong performance across disparity-centered metrics, with fewer large local correspondence errors around specular highlights, homogeneous tissue, and disparity discontinuities. Ablations confirm the contributions of both the latent correction trajectory and reliability-gated candidate contribution.

The model processes stereo pairs independently, and both candidates can fail in severely degraded regions. Future work will investigate spatio-temporal latent refinement for dynamic endoscopic scenes. Appendix XIII discusses geometric assumptions, domain scope, and deployment limitations. Broader impact is discussed in Appendix I.

Acknowledgment

This work is supported by National Natural Science Foundation of China (Grant No. 62306003) and Science and Technology Innovation Project of Anhui University Chuzhou Research Institute (Grant No.KJCX-202404)

Competing Interests

All authors declare no competing interests.

References

[1] Danail Stoyanov, Marco Visentini Scarzanella, Philip Pratt, and Guang-Zhong Yang. Real-time stereo reconstruction in robotically assisted minimally invasive surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 275–282. Springer, 2010.

[2] David Recasens, José Lamarca, José M Fácil, JM Martinez Montiel, and Javier Civera. Endo-depth-and-motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints. *IEEE Robotics and Automation Letters*, 6(4):7225–7232, 2021.

[3] Max Allan, Jonathan Mcleod, Congcong Wang, Jean Claude Rosenthal, Zhenglei Hu, Niklas Gard, Peter Eisert, Ke Xue Fu, Trevor Zeffiro, Wenyao Xia, et al. Stereo correspondence and reconstruction of endoscopic data challenge. *arXiv preprint arXiv:2101.01133*, 2021.

[4] Faisal Mahmood and Nicholas J Durr. Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy. *Medical image analysis*, 48:230–243, 2018.

[5] Patrice Monkam, Jing Wu, Wenkai Lu, Wenjun Shan, Hao Chen, and Yuhao Zhai. Easyspec: Automatic specular reflection detection and suppression from endoscopic images. *IEEE Transactions on Computational Imaging*, 7:1031–1043, 2021.

[6] Menglong Ye, Edward Johns, Ankur Handa, Lin Zhang, Philip Pratt, and Guang-Zhong Yang. Self-supervised siamese learning on stereo image pairs for depth estimation in robotic surgery. *arXiv preprint arXiv:1705.08260*, 2017.

[7] Shuwei Shao, Zhongcai Pei, Weihai Chen, Wentao Zhu, Xingming Wu, Dianmin Sun, and Baochang Zhang. Self-supervised monocular depth and ego-motion estimation in endoscopy: Appearance flow to the rescue. *Medical image analysis*, 77:102338, 2022.

[8] Bin Pu, Xingguo Lv, Jiewen Yang, Guannan He, Xingbo Dong, Yiqun Lin, Shengli Li, Ying Tan, Fei Liu, Ming Chen, Zhe Jin, Kenli Li, and Xiaomeng Li. Unsupervised domain adaptation for anatomical structure detection in ultrasound images. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 41204–41220. PMLR, 2024.

[9] Bin Pu, Jiewen Yang, Xingguo Lv, Xingbo Dong, Lei Zhao, Shengli Li, Kenli Li, and Xiaomeng Li. ToMo-UDA++: Unsupervised domain adaptation for anatomical structure detection using enhanced topology and morphology knowledge. *International Journal of Computer Vision*, 134, 2026. Article 230.

[10] Jiewen Yang, Yiqun Lin, Bin Pu, Jiarong Guo, Xiaowei Xu, and Xiaomeng Li. CardiacNet: Learning to reconstruct abnormalities for cardiac disease assessment from echocardiogram videos. In *Computer Vision – ECCV 2024*, volume 15081 of *Lecture Notes in Computer Science*, pages 293–311. Springer, 2024.

[11] Danail Stoyanov. Stereoscopic scene flow for robotic assisted minimally invasive surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 479–486. Springer, 2012.

[12] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2008.

[13] Bojian Li, Bo Liu, Miao Zhu, Xiaoyan Luo, and Fugen Zhou. Image intrinsic-based unsupervised monocular depth estimation in endoscopy. *IEEE Journal of Biomedical and Health Informatics*, 2024.

[14] Ying Zhou, Shiquan He, Hao Wang, Fan Huang, Mei Liu, Qiang Li, and Zhiwei Wang. Improved self-supervised monocular endoscopic depth estimation based on pose alignment-friendly dynamic view selection. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3005–3012. IEEE, 2024.

[15] Zhiwei Wang, Ying Zhou, Shiquan He, Ting Li, Fan Huang, Qiang Ding, Xinxia Feng, Mei Liu, and Qiang Li. Monopcc: Photometric-invariant cycle constraint for monocular depth estimation of endoscopic images. *Medical Image Analysis*, 102:103534, 2025.

[16] Beilei Cui, Mobarakol Islam, Long Bai, An Wang, and Hongliang Ren. Endodac: Efficient adapting foundation model for self-supervised depth estimation from any endoscopic camera. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 208–218. Springer, 2024.

[17] Liangjing Shao, Linxin Bai, Chenkang Du, and Xinrong Chen. Endomust: Monocular depth estimation for robotic endoscopy via end-to-end multi-step self-supervised training. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7739–7746. IEEE, 2025.

[18] Rongqi Wang, Rui Song, Jingang Zhang, and Yunfeng Nie. Leveraging spatiotemporal cues for self-supervised stereo depth estimation in endoscopic videos. *IEEE Transactions on Medical Imaging*, 2026.

[19] Bojian Li, Bo Liu, Xinning Yao, Jinghua Yue, and Fugen Zhou. Advancing depth anything model for unsupervised monocular depth estimation in endoscopy. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4355–4362. IEEE, 2025.

[20] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE international conference on computer vision*, pages 66–75, 2017.

[21] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5418, 2018.

[22] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3273–3282, 2019.

[23] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International conference on 3D vision (3DV)*, pages 218–227. IEEE, 2021.

[24] Jiewen Yang, Xingbo Dong, Liujun Liu, Chao Zhang, Jiajun Shen, and Dahai Yu. Recurring the transformer for video action recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14043–14053. IEEE, 2022.

[25] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1959–1968, 2020.

[26] Vladimir Tankovich, Christian Hane, Yinda Zhang, Adarsh Kowdle, Sean Fanello, and Sofien Bouaziz. Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14362–14372, 2021.

[27] Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5249–5260, 2025.

[28] Junda Cheng, Longliang Liu, Gangwei Xu, Xianqi Wang, Zhaoxing Zhang, Yong Deng, Jinliang Zang, Yurui Chen, Zhipeng Cai, and Xin Yang. Monster: Marry monodepth to stereo unleashes power. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6273–6282, 2025.

[29] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10371–10381, 2024.

[30] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

[31] PJ Eddie Edwards, Dimitris Psychogyios, Stefanie Speidel, Lena Maier-Hein, and Danail Stoyanov. Serv-ct: A disparity dataset from cone-beam ct for validation of endoscopic 3d reconstruction. *Medical image analysis*, 76:102302, 2022.

[32] Jiewen Yang, Yiqun Lin, Bin Pu, and Xiaomeng Li. Bidirectional recurrence for cardiac motion tracking with Gaussian process latent coding. In *Advances in Neural Information Processing Systems*, volume 37, pages 34800–34823, 2024.

[33] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

[34] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[35] Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[36] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[37] Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

[38] Gangwei Xu, Junda Cheng, Peng Guo, and Xin Yang. Attention concatenation volume for accurate and efficient stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12981–12990, 2022.

[39] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21919–21928, 2023.

[40] Xianqi Wang, Gangwei Xu, Hao Jia, and Xin Yang. Selective-stereo: Adaptive frequency information selection for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19701–19710, 2024.

[41] Dimitrios Psychogyios, Evangelos Mazomenos, Francisco Vasconcelos, and Danail Stoyanov. Msdesis: Multitask stereo disparity estimation and surgical instrument segmentation. *IEEE transactions on medical imaging*, 41(11):3218–3230, 2022.

[42] Hongkuan Shi, Zhiwei Wang, Ying Zhou, Dun Li, Xin Yang, and Qiang Li. Bidirectional semi-supervised dual-branch cnn for robust 3d reconstruction of stereo endoscopic images via adaptive cross and parallel supervisions. *IEEE Transactions on Medical Imaging*, 42(11):3269–3282, 2023.

[43] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Dynamicstereo: Consistent dynamic depth from stereo videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13229–13239, 2023.

[44] Yang Ding, Can Han, Sijia Du, Yaqi Wang, and Dahong Qian. Robust real-time endoscopic stereo matching under fuzzy tissue boundaries. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 392–406. Springer, 2025.

[45] Xu Wang, Jialang Xu, Shuai Zhang, Baoru Huang, Danail Stoyanov, and Evangelos B Mazomenos. Stereomamba: Real-time and robust intraoperative stereo disparity estimation via long-range spatial dependencies. *IEEE Robotics and Automation Letters*, 2025.

[46] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020.

[47] Ziyang Zheng, Jiewen Yang, Xinpeng Ding, Xiaowei Xu, and Xiaomeng Li. GL-Fusion: Global-local fusion network for multi-view echocardiogram video segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, volume 14223 of *Lecture Notes in Computer Science*, pages 78–88. Springer, 2023.

Technical Appendices and Supplementary Material

I Broader Impact

Accurate dense stereo disparity estimation can improve metric perception in computer-assisted endoscopy. Manifold-Stereo targets scenes where specular reflection, weak texture, deformation, occlusion, or domain shift degrade local stereo correspondence. By retaining calibrated stereo geometry as the metric backbone and using monocular predictions as reliability-conditioned contextual cues, the framework is compatible with stereo endoscopic platforms that already acquire rectified image pairs.

More reliable dense disparity, and the depth reconstructed from it, may improve anatomical surface reconstruction, augmented-reality visualization, robotic perception, quantitative measurement, and simulation-based training. In surgical robotics, locally robust correspondence estimates could stabilize scene understanding around tissue folds, instruments, and specular regions, where photometric matching often degrades. In clinical research, improved geometric perception may support more accurate 3D datasets, quantitative endoscopic video analysis, and benchmarking of downstream perception and navigation modules.

This work contributes to robust geometric perception for endoscopic intelligence. Future translation can extend the same geometry-guided refinement principle with uncertainty estimation, calibration monitoring, out-of-distribution detection, temporal consistency, and prospective multi-institution evaluation. These extensions would connect benchmark progress to practical systems while preserving clinician oversight and privacy-preserving data governance.

II Technical Scope of the Latent Residual Flow

The latent residual flow implements the correction component of Manifold-Stereo's reliability-gated latent-flow design. The stereo cost-volume encoder maps the geometric encoding volume to an image-aligned latent state Z_0, the residual decoder maps latent states to disparity corrections, and a conditional vector field generates a time-indexed path from Z_0 to a refined state Z_T. Here, "flow" denotes the time-one map of the learned vector field used for residual candidate refinement, not a distributional flow-matching objective. Refinement is supervised through decoded disparity predictions, while endpoint anchoring and path-energy regularization discourage large latent departures from the cost-volume embedding.

The geometric terminology is local and operational. The cost-volume latent state together with the residual decoder defines a local correction space, which is the family of disparity corrections producible around the current stereo evidence. This does not claim that endoscopic anatomy lies on a learned global manifold [35–37]. Instead, the construction provides a controlled parameterization for local residual corrections to the monocular candidate, and the geometric consistency gate determines the corrected candidate's contribution during recurrent stereo refinement.

Why use an ODE-style update. The continuous-time notation is used as a compact parameterization for repeated residual correction in the cost-volume latent field. In the implementation, Z_0 is not sampled noise; it is the spatial latent field encoded from the stereo cost volume. The vector field takes $[t, Z_t(p), C_{\text{field}}(p)]$ at each image location, and an 8-step RK4 solver repeatedly evaluates the same MLP vector field before a 3×3 residual decoder maps the endpoint to a disparity offset. The solver therefore increases the correction path length and numerical integration accuracy, but it does not add learnable parameters. This design differs from adding a deeper CNN residual head because the intermediate latent states are explicitly tied by a shared vector field and are directly regularized by endpoint anchoring and path energy. It also differs from generative flow matching because there is no distribution-transport objective, no likelihood, and no sampled target distribution.

The one-step ablation uses the same cost-volume latent encoder, conditioning projection, vector field, and residual decoder, but evaluates the vector field once instead of integrating the trajectory. This comparison tests repeated integration with shared parameters; the no-path-energy ablation tests trajectory regularization separately. We therefore interpret the gain from the full model as evidence that a regularized multi-step latent correction is useful for preparing the monocular candidate before reliability-gated stereo refinement, not as evidence for a global anatomical manifold or a generative flow model.

III Evaluation Metrics

We define the disparity metrics used in the main experiments and the calibrated Depth MAE reported in the SCARED comparison. To avoid ambiguity, we clarify that throughout the paper MAE (mm) refers to calibrated depth mean absolute error, not disparity MAE. Average disparity error is reported as EPE (px). Let Ω_v denote the valid evaluation pixels and $N = |\Omega_v|$. Let $\hat{d}_i$ and d_i be the predicted and ground-truth disparities at pixel i. Invalid pixels and pixels outside the dataset-specific evaluation mask are excluded from all averages. The disparity error is $e_i^d = |\hat{d}_i - d_i|$. We report

$$\text{EPE} = \frac{1}{N} \sum_{i \in \Omega_v} e_i^d, \tag{7}$$

and

$$\text{Bad-}t = \frac{100}{N} \left| \left\{ i \in \Omega_v : e_i^d > t \right\} \right|, \qquad t \in \{1, 2, 3, 5\}. \tag{8}$$

D1 denotes the dataset-specific stereo outlier rate and follows the corresponding benchmark threshold when provided by the evaluation protocol. Lower values are better for all disparity metrics.

For metric depth reporting, disparity is converted to depth by

$$\hat{y}_i = \frac{fb}{\max(\hat{d}_i, \epsilon)}, \qquad y_i = \frac{fb}{\max(d_i, \epsilon)}, \tag{9}$$

where f is the focal length, b is the stereo baseline, and ϵ prevents division by zero. Depth MAE in millimeters is

$$\text{MAE (mm)} = \frac{1}{N} \sum_{i \in \Omega_v} |\hat{y}_i - y_i|. \tag{10}$$

Lower values are better for MAE (mm).

IV Training Objective Details

We use recurrent sequence supervision for disparity prediction. Let Ω_v denote the valid disparity pixels after dataset masking and disparity-range filtering. The model first predicts an initial disparity D_{init} from the regularized cost volume, then produces a shared recurrent prediction sequence $\{D^k\}_{k=1}^K$ containing stereo-branch predictions and late-stage monocular-branch predictions inserted during stereo-monocular refinement. The sequence loss is

$$\mathcal{L}_{\text{seq}} = \frac{1}{|\Omega_v|} \sum_{p \in \Omega_v} \text{smooth}_{\ell_1} \left(D_{\text{init}}(p) - D^{\text{gt}}(p) \right) + \sum_{k=1}^{K} \omega_k \frac{1}{|\Omega_v|} \sum_{p \in \Omega_v} \left| D^k(p) - D^{\text{gt}}(p) \right|, \tag{11}$$

where the recurrent weights follow the RAFT/RAFT-Stereo-style recurrent training schedule [23, 46]. For $K > 1$,

$$\omega_k = \left(\gamma^{\frac{15}{(K-1)}} \right)^{K-k} = \gamma^{\frac{15(K-k)}{(K-1)}}, \qquad \gamma = 0.9. \tag{12}$$

For $K = 1$, we set $\omega_1 = 1$. The constant 15 is an empirical schedule-normalization factor inherited from this recurrent-loss convention. It keeps the total decay range approximately fixed as the number of recurrent predictions K changes. When $K = 16$, the weighting reduces to $\omega_k = \gamma^{K-k}$. For other $K > 1$, the exponent is rescaled to maintain a comparable relative weighting range across update budgets. Monocular-branch refinement is therefore supervised through the same recurrent prediction sequence rather than a separate independently weighted objective.

Per-pair stereo-scale alignment and geometry-aware weighting. For affine alignment of the refined monocular disparity, valid fitting pixels must have positive current stereo disparity, refined monocular disparity larger than 10^{-2}, and refined monocular disparity above its image-wise 20th percentile. The affine fitting is internal to the model and uses the current stereo prediction, not ground-truth disparity. The percentile filter removes near-zero monocular values that typically correspond to invalid or unstable estimates. This least-squares fitting uses no geometric-consistency threshold and no inference-time left-right consistency check.

The geometry-aware auxiliary term is implemented without a separate right-to-left disparity mask. Let $D_{\text{geo}} = \text{sg}(D^{\text{up}}_{\text{mono,ref}})$ be the detached upsampled monocular-branch disparity after latent and recurrent

refinement. Let $u(p) = \sqrt{\|f_L(p) - \mathrm{Warp}(f_R, D_{\text{stereo}})(p)\|_2^2 + \epsilon}$ denote the feature uncertainty induced by the current stereo disparity. We normalize and clip it as

$$\bar{u}(p) = \frac{\mathrm{clip}(u(p)/(\mathrm{mean}(u) + \epsilon), 0, 3)}{\max_q \mathrm{clip}(u(q)/(\mathrm{mean}(u) + \epsilon), 0, 3) + \epsilon}, \tag{13}$$

and combine it with the gate-derived weight $1 - \mathcal{G}_{\text{stereo}}$ as

$$w(p) = \frac{1}{2}\bar{u}(p) + \frac{1}{2}\left(1 - \mathcal{G}_{\text{stereo}}(p)\right). \tag{14}$$

The auxiliary consistency term is the weighted smooth-ℓ_1 agreement between D_{final} and D_{geo} over the training valid mask, as defined in the main Method section. The implementation also includes a small consistency regularizer proportional to $-\mathbb{E}_{p\in\Omega_v}[\mathcal{G}_{\text{stereo}}(p)]$. This term uses no swapped-pair forward pass, secondary right-disparity head, offline right-to-left pseudo-label, or threshold ϵ_{lr}.

V Mathematical Interpretation of Latent Candidate Refinement and Reliability-Gated Contribution

We give an operational interpretation of cost-volume-guided correction with reliability-gated candidate contribution. Conditional latent residual flow refines the monocular candidate in a local correction space, and reliability-weighted contribution control determines where the corrected candidate contributes to recurrent stereo-monocular updates. The formulation is local to the current stereo pair and cost volume. It does not require a global anatomical surface model. Instead, it specifies how the learned latent state and residual decoder parameterize a local family of disparity residuals.

Local correction space. Let $\Omega \subset \mathbb{Z}^2$ denote the image lattice at $\frac{1}{4}$ resolution and let $\mathcal{X} = \mathbb{R}^{|\Omega|}$ be the ambient vector space of discrete disparity maps. Let V_{geo} be the geometric encoding volume produced by the 3D aggregation network. The latent encoder defines

$$E_\theta : \mathcal{V} \to \mathcal{Z}, \qquad Z_0 = E_\theta(V_{\text{geo}}) \in \mathbb{R}^{L\times|\Omega|}, \tag{15}$$

where $L = 128$ and $Z_0(:, p)$ is the latent feature associated with pixel $p \in \Omega$. A residual decoder $\psi_\theta : \mathcal{Z} \to \mathcal{X}$ maps a latent field to a disparity correction. For a compact neighborhood $\mathcal{U}(Z_0) \subset \mathcal{Z}$, the decoder induces the local correction space

$$\mathcal{C}_\theta(D_{\text{mono}}, V_{\text{geo}}) = \{D_{\text{mono}} + \psi_\theta(Z) : Z \in \mathcal{U}(Z_0)\} \subset \mathcal{X}. \tag{16}$$

This local correction family is the operational object used by the method. The learned decoder is not required to satisfy a global rank condition, and the construction is not used as a global anatomical manifold model. For interpretation, a regularized decoder-induced quadratic form measures how latent displacements change decoded disparity corrections,

$$g_Z(U, V) = \langle J_{\psi_\theta}(Z)U,\ J_{\psi_\theta}(Z)V\rangle + \lambda\langle U, V\rangle, \qquad \lambda > 0, \tag{17}$$

which measures the sensitivity of decoded disparity corrections to latent displacements. The Tikhonov term keeps this local measure non-degenerate when the learned decoder is locally low-rank. The method therefore uses a stereo-conditioned local correction space for residual correction, not a global surface prior.

Conditional latent residual flow. Given a conditioning field C_{field} extracted from the matching descriptors, refinement is defined by the non-autonomous ODE

$$\frac{\mathrm{d}Z_t}{\mathrm{d}t} = v_\theta(t, Z_t, C_{\text{field}}), \qquad Z_{t=0} = Z_0, \quad t \in [0, 1], \tag{18}$$

where v_θ is the MLP vector field used in the implementation. The refined latent state is $Z_T = \Phi_T(Z_0, C_{\text{field}})$, where Φ_t denotes the flow map induced by the ODE. The refined monocular candidate is

$$\tilde{D}_{\text{mono}} = D_{\text{mono}} + \psi_\theta(Z_T). \tag{19}$$

The vector field therefore does not directly predict the final disparity. It predicts a path in the local correction space, and the residual decoder converts the endpoint into a disparity update.

Stable local latent trajectory. Assume that $v_\theta(t, Z, C)$ is continuous in t and locally Lipschitz in Z for fixed C. Let $Z_0 \in \mathcal{U}$, where $\mathcal{U} \subset \mathcal{Z}$ is a compact latent neighborhood whose distance from Z_0 to the neighborhood boundary is $r > 0$. If

$$\int_0^1 \|v_\theta(t, Z_t, C)\|_2 \, \mathrm{d}t < r, \tag{20}$$

then the ODE has a unique solution on $[0, 1]$, the trajectory remains inside $\mathcal{U}$, and the decoded path $\psi_\theta(Z_t)$ remains inside the decoder image $\psi_\theta(\mathcal{U})$ of that local latent neighborhood.

Justification. Existence and uniqueness follow from the Picard-Lindelof theorem under the stated continuity and Lipschitz conditions. The integral bound limits the latent path length, so the trajectory cannot cross the boundary of $\mathcal{U}$ within unit time. This sufficient local condition supports the bounded-residual interpretation. It is not a global guarantee for all inputs or surgical scenes.

Training objective as local path regularization. The network does not require a manually annotated latent endpoint. Dense disparity supervision constrains decoded stereo and monocular predictions, while the latent-flow regularizer keeps the trajectory near the cost-volume-induced local correction neighborhood. Let

$$\bar{Z} = \mathrm{sg}(Z_0) \tag{21}$$

denote a stop-gradient reference state, where $Z_0 = E_\theta(V_{\text{geo}})$ and $\mathrm{sg}(\cdot)$ stops gradients through this reference. For the discrete integration states $\{Z_n\}_{n=0}^N$ with $Z_N = Z_T$, the training objective uses

$$\mathcal{L}_{\text{flow}} = \|Z_T - \bar{Z}\|_2^2 + \alpha \sum_{n=0}^{N-1} \|Z_{n+1} - Z_n\|_2^2, \tag{22}$$

which anchors the endpoint to the local embedding center while penalizing unnecessarily long trajectories. The first term discourages large endpoint displacement, and the second controls path energy. Non-trivial refinement remains possible because Z_T is decoded by ψ_θ and receives gradients from supervised disparity losses. The regularizer constrains the latent update without forcing the decoded correction to zero.

Relation to local energy regularization. For a differentiable latent path $\gamma : [0, 1] \to \mathcal{Z}$, the decoder-induced quadratic form gives the local energy

$$\mathcal{E}_g(\gamma) = \int_0^1 g_{\gamma(t)}(\dot{\gamma}(t), \dot{\gamma}(t)) \, \mathrm{d}t. \tag{23}$$

The exact metric tensor g is not computed explicitly in the network. Instead, the implementation uses the discrete latent path penalty

$$\mathcal{E}_{\text{lat}} = \sum_{n=0}^{N-1} \|Z_{n+1} - Z_n\|_2^2. \tag{24}$$

This penalty is a local surrogate when the decoder-induced quadratic form is bounded in the latent neighborhood. Specifically, if $mI \preceq G(Z) \preceq MI$ for the matrix $G(Z)$ in a local neighborhood, then for small integration steps

$$m \sum_{n=0}^{N-1} \|\Delta Z_n\|_2^2 \le \sum_{n=0}^{N-1} \Delta Z_n^{\mathsf{T}} G(Z_n) \Delta Z_n \le M \sum_{n=0}^{N-1} \|\Delta Z_n\|_2^2, \tag{25}$$

where $\Delta Z_n = Z_{n+1} - Z_n$. Under this bounded-sensitivity condition, controlling the Euclidean latent trajectory length also controls the decoder-induced local energy up to constants. This motivates the compact path regularizer used in the implementation, without requiring explicit estimation of a full metric tensor.

Reliability-weighted contribution. Let the feature residual under a disparity hypothesis D be

$$r(D, p) = f_L(p) - \mathrm{Warp}(f_R, D)(p). \tag{26}$$

The geometric consistency score used in the model,

$$\mathcal{G}_{\text{stereo}}(p) = \exp\left(-\frac{\|r(D^k_{\text{stereo}}, p)\|_2^2}{C\tau}\right), \tag{27}$$

is used as a soft reliability weight for the current stereo correspondence. For a local representation ξ with stereo and monocular candidates ξ_{stereo} and ξ_{mono}, consider the quadratic surrogate

$$\min_{\xi} \ \mathcal{G}_{\text{stereo}}\|\xi - \xi_{\text{stereo}}\|_2^2 + (1 - \mathcal{G}_{\text{stereo}})\|\xi - \xi_{\text{mono}}\|_2^2. \tag{28}$$

The minimizer is

$$\xi^\star = \mathcal{G}_{\text{stereo}}\xi_{\text{stereo}} + (1 - \mathcal{G}_{\text{stereo}})\xi_{\text{mono}}. \tag{29}$$

The gating rule in Manifold-Stereo therefore has a least-squares interpretation. It is the minimizer of this local weighted quadratic surrogate. High-consistency regions remain closer to the stereo representation, while low-consistency regions assign greater weight to the corrected monocular representation.

Consequence for the full refinement model. The latent residual flow and reliability gate form a single cost-volume-guided stereo-monocular refinement process. The flow defines a stereo-conditioned correction trajectory for the monocular candidate, and the gate determines the corrected candidate's contribution to the recurrent stereo state. This candidate-contribution view is well suited to endoscopy. Ambiguous tissue regions can benefit from monocular context, while well-matched regions remain dominated by calibrated stereo correspondence.

VI Domain Applicability of Reliability-Gated Candidate Contribution

This section clarifies the domain assumptions behind Manifold-Stereo. The method is not a proof that reliability-gated latent refinement cannot be used for natural-image stereo. Instead, it specifies conditions under which the design is expected to be especially effective, and explains why calibrated endoscopic stereo often satisfies these conditions more directly than broad outdoor or driving scenes.

Figure I illustrates this domain contrast. Outdoor driving scenes contain heterogeneous object categories, thin structures, independently positioned depth layers, large occlusions, reflective surfaces, and strong scale variation. Their disparity maps are therefore commonly piecewise smooth, with frequent discontinuities at object and occlusion boundaries. Endoscopic stereo instead often observes connected anatomical tissue surfaces whose disparity varies more coherently over local regions, while correspondence failures are concentrated around specular reflection, smoke, wet tissue, fluid, instruments, folds, and contact boundaries. This structure motivates a reliability-gated contribution rule: calibrated stereo should remain dominant where left-right alignment is reliable, while the corrected monocular candidate can support localized ambiguous tissue regions.

Error decomposition for gated contribution. Let D_{s} denote the current stereo estimate and let $\tilde{D}_{\text{m}}$ denote the corrected monocular candidate after stereo-scale alignment. For a pixel p, define the local errors

$$e_{\text{s}}(p) = D_{\text{s}}(p) - D^{\text{gt}}(p), \qquad e_{\text{m}}(p) = \tilde{D}_{\text{m}}(p) - D^{\text{gt}}(p). \tag{30}$$

For an illustrative scalar mixture (the implemented gate combines update features), define

$$D_{\text{g}}(p) = g(p)D_{\text{s}}(p) + (1 - g(p))\,\tilde{D}_{\text{m}}(p), \qquad g(p) \in [0, 1], \tag{31}$$

with error

$$e_{\text{g}}(p) = g(p)e_{\text{s}}(p) + (1 - g(p))\,e_{\text{m}}(p). \tag{32}$$

Assuming locally zero-mean errors with variances $\sigma_{\text{s}}^2(p)$ and $\sigma_{\text{m}}^2(p)$ and correlation $\rho(p)$, the expected squared error is

$$\mathbb{E}\left[e_{\text{g}}^2(p)\right] = g^2\sigma_{\text{s}}^2 + (1 - g)^2\sigma_{\text{m}}^2 + 2g(1 - g)\rho\sigma_{\text{s}}\sigma_{\text{m}}. \tag{33}$$

When the denominator is positive, the unconstrained variance-minimizing weight is

$$g^\star(p) = \frac{\sigma_{\text{m}}^2(p) - \rho(p)\sigma_{\text{s}}(p)\sigma_{\text{m}}(p)}{\sigma_{\text{s}}^2(p) + \sigma_{\text{m}}^2(p) - 2\rho(p)\sigma_{\text{s}}(p)\sigma_{\text{m}}(p)}. \tag{34}$$

For the constrained mixture, this solution is clipped to $[0, 1]$; if the denominator is zero, the two errors coincide almost surely and the weight does not change the mixture error. The expression illustrates the desired behavior: when stereo error variance is much smaller, the constrained weight approaches one; when the corrected monocular candidate is more stable, it decreases. Manifold-Stereo does not estimate these variances explicitly or claim to recover this optimum. It uses left-right feature consistency as an empirical proxy for stereo reliability.

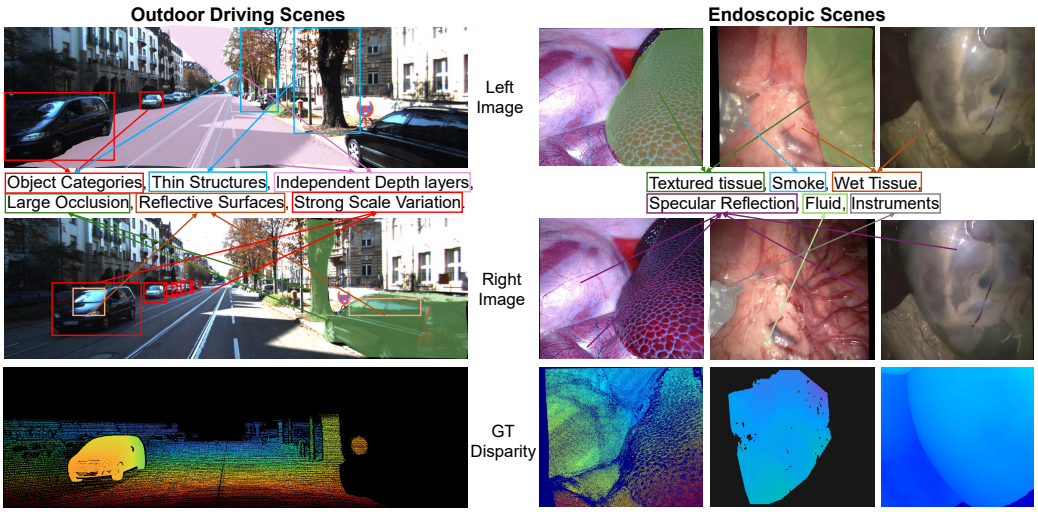

Figure I: **Domain contrast between natural-image stereo and endoscopic stereo.** Outdoor driving scenes often contain heterogeneous object categories, thin structures, independently positioned depth layers, large occlusions, reflective surfaces, and strong scale variation, producing disparity fields that are piecewise smooth with frequent object- and occlusion-boundary discontinuities. Endoscopic scenes more often observe connected tissue surfaces with locally coherent disparity, while local correspondence ambiguity arises from specular reflection, smoke, wet tissue, fluid, instruments, folds, and contact boundaries. This contrast motivates reliability-gated candidate contribution for calibrated endoscopic stereo, where monocular context is introduced selectively in locally ambiguous tissue regions while stereo correspondence remains the metric anchor in reliable regions.

Endoscopic scene structure and local-coherence assumptions. Endoscopic scenes often contain connected tissue regions whose disparity varies coherently over local neighborhoods, except near folds, tools, occlusions, and contact boundaries. Let $\Omega = \Omega_{\text{rel}} \cup \Omega_{\text{amb}}$, where Ω_{rel} contains pixels with reliable left-right support and Ω_{amb} contains locally ambiguous pixels caused by weak texture, specular appearance, fluid, smoke, instruments, or occlusion. In many endoscopic frames, Ω_{amb} is embedded in or adjacent to the same anatomical surface as Ω_{rel}. A monocular candidate can therefore provide coherent contextual estimates over ambiguous tissue patches, while calibrated stereo still provides the metric anchor in neighboring reliable regions.

The latent residual flow is designed for this regime. The cost-volume-induced latent state anchors corrections to the stereo evidence of the current pair, and the path regularizer discourages large departures from this local embedding. The gate then prevents the corrected monocular candidate from contributing uniformly across the image. As a result, the method matches the common endoscopic structure of broad tissue surfaces with localized correspondence failures.

Comparison with natural-image stereo. Natural-image stereo, such as outdoor driving scenes, can contain many object categories, thin structures, independently positioned depth layers, large occlusion boundaries, reflective surfaces, and strong scale variation. The resulting disparity field is often piecewise smooth rather than locally coherent over broad connected tissue-like regions, and it can have higher boundary density and less shared surface support across neighboring pixels. In this setting, the assumption that an ambiguous region can be corrected by a locally coherent monocular candidate anchored to nearby stereo evidence is weaker. The reliability gate can still be useful, but the monocular candidate may require object-aware or scene-specific priors, and the correction space may need to model more heterogeneous geometry.

Thus, Manifold-Stereo is best interpreted as a method targeted to calibrated endoscopic stereo, where local stereo reliability varies sharply but the underlying tissue geometry often remains regionally coherent. Applying the same design to natural-image stereo would require separate validation and may benefit from additional mechanisms for object boundaries, semantic diversity, and broader scene layout.

VII Operational Details of Candidate Construction and Output Recovery

We specify the supporting operations that construct the inputs and outputs of the reliability-gated latent-flow framework. These operations are not separate conceptual contributions. The stereo stem preserves high-frequency correspondence cues for metric geometry, the pre-trained monocular encoder provides larger-context anatomical features and an uncalibrated disparity-like candidate, and the output decoder restores full-resolution disparity after cost-volume-guided recurrent refinement.

Stereo correspondence branch. The stereo branch is a four-level convolutional stem with instance normalization. For an input image, the stem produces features at $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, and $\frac{1}{16}$ resolution as

$$3 \to 32 \to 48 \to 96 \to 192. \tag{35}$$

Each level uses a stride-2 `BasicConv_IN` block with a 3×3 kernel, followed by a 3×3 convolution, instance normalization, and ReLU activation. The first block maps RGB input to 32 channels at $\frac{1}{2}$ resolution. Subsequent blocks map $32 \to 48$, $48 \to 96$, and $96 \to 192$ channels. Instance normalization is used because endoscopic frames often exhibit procedure-specific illumination and color statistics. Normalizing each sample reduces appearance sensitivity while retaining local spatial structure for stereo matching.

The $\frac{1}{4}$ stereo stem feature is concatenated with the transferred monocular feature at the same scale. This produces a 96-channel descriptor input, which is processed by a 3×3 `BasicConv_IN` layer and a 1×1 descriptor projection, yielding left and right matching descriptors $f_L, f_R \in \mathbb{R}^{B \times 96 \times H/4 \times W/4}$. The right branch is used symmetrically for matching descriptors, while the deeper left stem features are also used to construct recurrent context features.

Monocular feature transfer. The monocular branch uses the pre-trained Depth Anything V2 encoder and decoder [30] to produce multi-scale features. The encoder and monocular depth decoder are frozen, while a feature-transfer operation adapts the decoder features for stereo matching. The transfer operation maps the monocular features into four scales with channel dimensions 48, 64, 192, and 160. It first processes the coarsest feature with a 3×3 convolution to produce a 160-channel $\frac{1}{32}$ representation, then progressively upsamples it by transposed convolutions and combines it with finer features using 5×5 convolutions and 1×1 residual projections. This top-down construction provides global monocular context to finer matching descriptors without discarding local details needed for stereo correspondence.

Group-wise cost volume. Given the left and right descriptors f_L and f_R, the implementation constructs a group-wise correlation volume with $G = 8$ groups and maximum disparity $D_{\max}/4$. For each disparity d, the right descriptor is shifted by d pixels and correlated with the left descriptor within each channel group. The output volume has shape

$$V_{\text{gwc}} \in \mathbb{R}^{B \times 8 \times \frac{D_{\max}}{4} \times \frac{H}{4} \times \frac{W}{4}}. \tag{36}$$

The group-wise design preserves multiple matching subspaces instead of collapsing all channels into a single correlation score, allowing different feature groups to encode tissue texture, boundary structure, and illumination variation.

3D aggregation network. The cost volume is first processed by a $3 \times 3 \times 3$ 3D convolution that preserves the 8-channel dimensionality. A feature attention module then uses the $\frac{1}{4}$ left feature to modulate the volume. The main cost aggregation module is a 3D hourglass network with three downsampling stages as

$$8 \to 16 \to 32 \to 48. \tag{37}$$

Each downsampling stage consists of a stride-2 3D convolution followed by a stride-1 3D convolution. The intermediate volumes are modulated by monocular feature attention at $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$ scales, using feature channels 64, 192, and 160, respectively. The decoder uses transposed 3D convolutions to restore the volume resolution. Skip connections concatenate decoder features with encoder features, followed by $1 \times 1 \times 1$ and $3 \times 3 \times 3$ aggregation convolutions. Additional feature attention is applied during upsampling at the $\frac{1}{16}$ and $\frac{1}{8}$ scales. The final transposed 3D convolution maps the aggregated representation back to 8 channels, producing the geometric encoding volume V_{geo}.

The disparity classifier is a $3 \times 3 \times 3$ convolution from 8 channels to 1 channel. A softmax over the disparity dimension produces a probability volume, and disparity regression computes the initial

disparity as the expectation over disparity candidates. This implementation keeps the cost regularizer compact while allowing the network to reason jointly over spatial neighborhoods and disparity hypotheses.

Recurrent correlation sampler. The recurrent update block uses a combined geometric encoding sampler. It takes the left descriptor, right descriptor, and V_{geo} as input. In addition to the learned geometric encoding volume, it constructs an all-pairs descriptor correlation along the rectified scanline. Both V_{geo} and the descriptor correlation are reshaped per left-image location and average-pooled along the disparity or right-coordinate dimension to form a small pyramid with `corr_levels` levels. At each recurrent iteration, the sampler bilinearly samples a local one-dimensional window of radius `corr_radius` around the current disparity at every $\frac{1}{4}$-resolution pixel. The sampled features are concatenated and passed to the $\frac{1}{4}$-resolution motion encoder. The $\frac{1}{8}$ and $\frac{1}{16}$ recurrent states are updated through pooled or interpolated hidden-state exchange rather than through separate spatial correlation volumes.

Spatial latent field and residual-flow refinement. The spatial latent field is obtained from V_{geo} using a $3 \times 3 \times 3$ convolution from 8 to 128 channels, ReLU activation, and adaptive average pooling only along the disparity dimension. The resulting tensor has shape

$$Z_0 \in \mathbb{R}^{B \times 128 \times \frac{H}{4} \times \frac{W}{4}}. \tag{38}$$

The flow conditioning field is generated from the unwarped right matching descriptor by a 1×1 convolution from 96 to 128 channels. This descriptor is used only as right-view matching context for the vector field, not as a point-wise scene-aligned feature concatenated with the left-view latent field. The latent field Z_0 is already encoded from the disparity-indexed cost volume, and explicit left-right alignment is handled by the group-wise cost volume and the recurrent geometric consistency gate. For each spatial location p, the MLP input is the concatenation $[t, Z_t(p), C_{\text{field}}(p)]$, where t is the scalar integration time. The conditional vector field is parameterized by a three-layer MLP with hidden dimension 256 and SiLU activations as

$$(1 + 128 + 128) \to 256 \to 256 \to 128. \tag{39}$$

The implementation flattens spatial locations into the batch dimension and integrates the vector field for 8 steps using RK4 from `torchdiffeq`. A 3×3 convolution maps the refined latent field from 128 channels to a one-channel $\frac{1}{4}$-resolution disparity offset, which updates the monocular disparity candidate before reliability-gated recurrent refinement.

Full-resolution output recovery. The final full-resolution prediction uses an output residual decoder. It takes $D^{\text{up}}_{\text{mono,ref}}$, $D^{\text{up}}_{\text{stereo}}$, I_L, and I_R as input, where $D^{\text{up}}_{\text{mono,ref}}$ is the upsampled monocular-branch disparity after latent and recurrent refinement. The right image is first warped twice, using $D^{\text{up}}_{\text{mono,ref}}$ and $D^{\text{up}}_{\text{stereo}}$, to obtain photometric residuals

$$e_{\text{mono}} = \text{Warp}(I_R, D^{\text{up}}_{\text{mono,ref}}) - I_L, \qquad e_{\text{stereo}} = \text{Warp}(I_R, D^{\text{up}}_{\text{stereo}}) - I_L. \tag{40}$$

Each residual is concatenated with the left image and projected by a 3×3 convolution from 6 to 16 channels. The monocular and stereo disparities are separately projected by 3×3 convolutions from 1 to 16 channels. These four 16-channel tensors are concatenated into a 64-channel representation, followed by a dilated 3×3 convolution with dilation 2 that maps the representation to 32 channels.

The 32-channel representation is refined by a two-pass encoder–decoder block. Each encoder pass uses four stride-2 convolutional stages with channel dimensions $32 \to 48 \to 64 \to 96 \to 128$. The decoder uses transposed convolutions with skip connections to restore full resolution. This block preserves local image structure while allowing the decoder to aggregate broader spatial context around boundaries and residual matching errors. Let F denote the resulting encoder–decoder feature. The decoder then applies global-local modulation to F [cf. global-local fusion in 47]. The global modulation feature G is obtained by adaptive average pooling followed by two 1×1 convolutions, and the local modulation mask M is obtained from the same feature F by three 3×3 convolutions. The refined feature is

$$F_{\text{dec}} = (1 - M) \odot G + M \odot F, \tag{41}$$

where G is broadcast spatially. A final 3×3 convolution maps F_{dec} to the one-channel refinement value R_{dec}, and the final disparity is $D_{\text{final}} = D^{\text{up}}_{\text{stereo}} + R_{\text{dec}}$, as described in the main Method section. This decoder is an output-resolution recovery module applied after the last cost-volume-guided recurrent update, rather than an additional reliability mechanism.

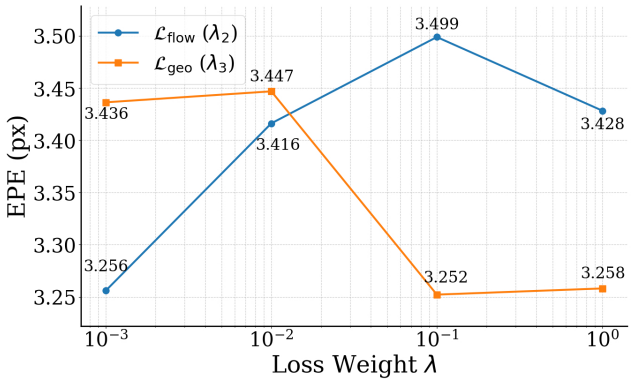

Figure II: Ablation of loss weights λ_2 and λ_3 for $\mathcal{L}_{\text{flow}}$ and $\mathcal{L}_{\text{geo}}$ in the training objective.

VIII Additional Ablation Details

The main paper reports ablations isolating the principal factors in cost-volume-guided correction with reliability-gated candidate contribution. Here, we provide additional sensitivity analyses for the recurrent update budget, geometry-aware training signal, and loss weights. These settings support reproducibility and complement the architectural comparisons in Tables 4–6.

Table I: Ablation of geometry-aware training signal on SCARED [3].

Training signal variant	**MAE (mm)**↓	**EPE (px)**↓
Full	**2.53**	**3.05**
No $\mathcal{L}_{\text{geo}}$	2.78	3.43

Table I evaluates the auxiliary geometry-aware loss while keeping recurrent sequence supervision unchanged. Removing $\mathcal{L}_{\text{geo}}$ increases both MAE and EPE, indicating that the auxiliary agreement term complements dense recurrent supervision by encouraging consistency with the detached monocular-branch geometric candidate in locally uncertain regions.

Table II: Ablation study on the number of iterations for scheduled recurrent stereo-monocular refinement on SCARED [3].

Iterations	**MAE (mm)**↓	**EPE (px)**↓
1	2.66	3.34
2	2.54	3.18
4	2.53	3.16
8	**2.53**	**3.05**
16	2.53	3.17
32	2.55	3.20

Table II studies the recurrent update budget under scheduled refinement. The model first performs stereo-only recurrent updates and then switches, after the $K - 8$ scheduling threshold, to stereo-monocular refinement after stereo-scale alignment with geometric consistency gating. Settings with 1–4 total recurrent iterations omit the full warm-up and late-stage reliability-gated candidate contribution stage, leading to higher MAE and EPE. With 8 iterations, the model reaches the minimum budget that enables an initial stereo update followed by scheduled gated refinement. Increasing the budget to 16 or 32 iterations adds earlier stereo-only updates but yields little additional gain, indicating that reliability-gated candidate contribution is more important than update count alone.

Figure II shows performance sensitivity as the weights of latent residual-flow regularization and geometric consistency vary independently. The best setting on the two curves occurs at $\lambda_2 = 10^{-3}$ and $\lambda_3 = 10^{-1}$, matching the default loss in Section 3.5. Larger λ_2 over-constrains the latent trajectory,

while insufficient λ_3 weakens the separation between reliable stereo regions and occluded or specular regions. This behavior matches the training design. Stereo geometry guides reliable correspondence, and latent residual refinement adjusts uncertain regions without suppressing image-specific matching information.

IX Roles of Monocular Features and the Disparity Candidate

The monocular branch supplies both transferred features and an explicit disparity candidate. These enter different parts of the computation (Table III). The initial latent state is encoded from the monocular-attended stereo cost volume:

$$Z_0 = E_\theta(V_{\text{geo}}), \qquad Z_T = \Phi_T(Z_0, C_{\text{field}}), \qquad \widetilde{D}_{\text{mono}} = D_{\text{mono}} + \psi_\theta(Z_T). \tag{42}$$

Transferred features help construct the matching representation, whereas the one-channel candidate supplies the base disparity estimate. The latent state is therefore not an explicit replacement for that candidate. The frozen monocular encoder and decoder produce features and a candidate for the same input, while stereo cost-volume geometry and matching descriptors condition the learned correction. Disparity supervision of the resulting predictions determines the required residual magnitude.

Table III: Distinct computational roles of monocular context, the latent state, and the disparity candidate.

Quantity	**Source / representation**	**Role**
Monocular features	Frozen monocular branch; transferred multi-scale features	Modulate matching descriptors and cost aggregation
Z_0	Monocular-attended stereo cost volume; 128-channel spatial field	Initialize the correction trajectory from pair-specific matching geometry
D_{mono}	Frozen monocular decoder; one-channel disparity-like map	Supply the base estimate to be corrected
$\psi_\theta(Z_T)$	Conditional latent residual flow and residual decoder; one-channel offset	Correct the base using stereo-conditioned latent refinement

Residual correction and the final prediction. The monocular decoder produces a dense proposal with unknown metric scale and offset. It is corrected, aligned to the current stereo scale, and introduced through geometric gating before the final recurrent prediction. Describing the refinement as residual correction distinguishes its output parameterization from a head trained to generate an absolute replacement disparity. This parameterization and its regularization provide an inductive bias; the additive expression alone is not a hard bound on the decoded residual. The term *latent candidate correction* refers to the same conditional latent residual-flow module, not an additional component.

X Ablation Scope and Component Controls

All ablations use the SCARED training and evaluation protocol in the main paper. We report the safeguards applied to the monocular proposal, the branch configurations, and the latent-refinement design separately. MAE denotes depth error in millimeters; EPE denotes disparity error in pixels. Lower values are better.

Safeguards for the monocular proposal. Table IV shows the EPE increase when spatial gating, multi-step correction, or stereo-scale alignment is removed or simplified, with the monocular candidate retained.

Monocular branch configurations. Table V specifies which operations are active. *Latent candidate correction* denotes the conditional latent residual flow. Stereo-only, formerly *No Monocular Feat.*, removes both the features and candidate, bypassing correction, alignment, and monocular contribution. The full-versus-stereo-only difference therefore measures the complete pathway, not the candidate alone.

Disabling the gated contribution removes its spatial reliability weighting; the monocular candidate remains present in the no-geometric-gate setting.

Table IV: Safeguards for the monocular proposal on SCARED.

Configuration	Removed / simplified component	EPE (px)↓	Increase over full
Full model	None	**3.05**	**0.00**
No geometric gate	Spatial control of monocular contribution	3.45	+0.40
One-step correction	Multi-step residual trajectory	3.55	+0.50
No stereo-scale alignment	Scale and offset alignment of the monocular candidate	4.32	+1.27

Table V: Monocular branch configurations on SCARED.

Configuration	Monocular feature transfer	Monocular disparity candidate	Latent candidate correction	Monocular gated contribution	MAE (mm)↓	EPE (px)↓
Full model	Enabled	Present	Enabled	Enabled	**2.53**	**3.05**
Stereo-only (No Monocular Branch)	Disabled	Absent	Bypassed	Absent	3.11	3.90
No correction space	Enabled	Present	Disabled	Enabled	2.60	3.27
No geometric gate	Enabled	Present	Enabled	Disabled	2.77	3.45

Latent-refinement design. Table VI isolates the multi-step trajectory and path-energy loss while retaining the monocular candidate.

Table VI: Latent-refinement design on SCARED.

Refinement	MAE (mm)↓	EPE (px)↓
Full latent refinement	**2.53**	**3.05**
One-step update	2.82	3.55
Without path-energy loss	3.09	3.84

The one-step variant retains the same encoder, conditioning projection, vector field, and decoder, but evaluates the vector field once. Its 0.50 px higher EPE supports multi-step integration without additional learnable parameters. Removing the path-energy loss leaves the forward architecture unchanged and increases EPE by 0.79 px.

XI Reliability-Gate Rationale and Diagnostics

Feature consistency versus cost-volume confidence. Cost-volume confidence summarizes ambiguity across discrete disparity hypotheses. In contrast, the feature residual tests left-right alignment at the current continuous disparity and is recomputed during refinement. The exponential score is a smooth, monotone mapping of nonnegative squared residual to $(0, 1]$: zero residual gives unit weight, and increasing residual decreases stereo contribution. It is a deterministic similarity kernel, not an assumed probability law of stereo correctness. The complementary weight allocates contribution inside the learned updater, whose inputs still include stereo correlation and state features. It does not imply that the monocular estimate must be accurate wherever stereo support is weak.

Gate diagnostic. Across the ten SCARED test keyframes, the top 20% of pixels by monocular weight have average stereo/monocular weights of approximately $0.14/0.86$. Inspected gates cluster around highlights, occlusion or depth boundaries, and photometric mismatch. This describes contribution allocation, not independent accuracy or probability calibration. The uniform-mixing control in Table 5 provides complementary evidence: replacing the spatial gate increases EPE from 3.05 to 3.30 px.

Challenging-region interpretation. Full-frame disparity plots can hide local changes because organ-scale depth gradients dominate their color range. Weak-texture tissue provides few discriminative matches; highlights and wet surfaces violate appearance consistency; smoke reduces local

contrast; and occlusions, tool contact, and deforming boundaries may introduce unmatched pixels or abrupt depth changes. Figures 3, 4, and 5 should be read at these local structures. The gate diagnostic supports selective contribution in ambiguous regions, while aggregate errors quantify prediction quality. Neither the diagnostic nor the full-frame views establish improvement in every degraded region; simultaneous candidate failures remain possible.

XII Backbone Choice and Evaluation Scope

The default MobileNetV2 stereo backbone extracts local matching cues, while the frozen ViT monocular branch supplies pretrained context. Table VII compares four stereo backbones within the refinement framework. MobileNetV2 gives the lowest EPE (3.05 px), while all four settings obtain MAE between 2.53 and 2.58 mm and EPE between 3.05 and 3.22 px. These results support the design across both convolutional and Transformer stereo backbones.

Table VII: Stereo-backbone ablation on SCARED.

Stereo backbone	**MAE (mm)**↓	**EPE (px)**↓
MobileNetV2 (Ours)	**2.53**	**3.05**
ResNet18	2.55	3.14
ConvNeXt	**2.53**	3.22
Swin-Tiny	2.58	3.21

All reproduced comparison methods use the same SCARED split and training-validation protocol, with SCARED-selected checkpoints evaluated directly on SERV-CT and Hamlyn. No target-dataset fine-tuning is performed. Differences in the method lists reflect input applicability: DynamicStereo requires sequences, so it is omitted on static SERV-CT pairs and unavailable for the single-frame SCARED K4 cases. Stereo-scale alignment is an internal operation fitted to the current stereo prediction, not target-label supervision. The natural-scene examples in Appendix VI illustrate domain structure; quantitative generalization claims here concern the evaluated datasets.

XIII Limitations and Scope

Local geometry. The manifold terminology describes a local family of decoded corrections around the current cost-volume state. No global constant-rank or anatomical-surface assumption is verified. The trajectory analysis requires the stated local regularity and bounded-sensitivity conditions, and the implemented loss is a soft regularizer rather than a hard projection into a certified neighborhood.

Shared failure modes. Feature consistency is an empirical proxy for stereo reliability. Both candidates may fail under severe specularity, occlusion, low contrast, or domain shift, and complementary weighting cannot by itself detect this shared failure. Stereo-scale alignment depends on the current stereo estimate and valid fitting pixels. These dependencies motivate future work on explicit uncertainty and failure detection.

Experimental scope. The gains are aggregate results. They do not imply best performance on every SCARED keyframe or every Hamlyn thresholded metric. Local smoothness assumptions may be less appropriate for natural scenes with many independently positioned objects and thin structures; transferring the method requires separate evaluation. The gate diagnostic characterizes contribution allocation, not calibrated uncertainty.

Temporal and clinical scope. The model processes frames independently and does not enforce temporal consistency. The evaluation concerns disparity estimation, not prospective clinical outcomes or validated surgical deployment. Future work should evaluate temporal stability, acquisition variability, and downstream reconstruction performance.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Please see our Abstract and Introduction Sections.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please see our Conclusion section and the Broader Impact in appendix section I.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We have provided the full set of assumptions and a complete (and correct) proof for all necessary parts.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have provided all information needed in our experimental section and appendix. (Please see section 3 and our appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will release all our code publicly upon the paper being accepted.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: Yes

   Justification: We have specified all the training and test details in our Result section. (Please see section 4)

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: the significance of the difference in disparity estimation results is validated by the T-test, our results are with $P \leq 0.05$, which denotes statistically significant.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please see section 4

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please see the section I in our Appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite these works and comply with their relevant agreements.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: [Yes]

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.