2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

CaLoRA-Stereo: Robust Stereo Endoscopic Depth Estimation Network via Camera-Aware LoRA and Dual-View Geometry

Shixing Ma[1], Shuwei Shao[1], Zhaoxi Lin[2], Xinzhe Du[1], Rui Song[1], Yibin Li[1], Max Q.-H. Meng[3], and Zhe Min[†1,4]

***Abstract*— Stereo depth estimation has drawn widespread attention from the robotics community due to its broad applications such as 3D reconstruction. Recently, stereo matching foundation models have made significant progress by being trained on the large-scale datasets containing natural images. However, directly leveraging these pretrained large models to minimally invasive surgery still remains challenging due to domain shifts in aspects of specular highlights and low-texture tissue. In this paper, we propose a parameter-efficient adaptation framework to address this gap. Specifically, we introduce Camera-Aware LoRA for fine-tuning Foundation-Stereo, using a camera-aware scaling gate computed from focal length and baseline to address intraoperative intrinsics drift arising from instrument self-heating and other thermal effects. We further develop a geometric consistency constraint and a spectral alignment regularizer that enforce cross-view depth agreement. Extensive experiments on the SCARED and Hamlyn datasets indicate that the proposed method achieves state-of-the-art performance. Notably, CaLoRA is easy to integrate into standard fine-tuning pipelines, requiring no backbone changes and only a small number of trainable parameters.**

I. Introduction

In minimally invasive surgery, stereo endoscopes provide two synchronized views of the operative field (Fig. 1), enabling depth perception that is crucial for navigation [1], [2], autonomous robotic control [3], [4], augmented reality [5]–[7], and digital twins [8]. Stereo matching computes disparity via epipolar correspondence and maps it to metric depth using focal length and baseline of camera [9]. Accurate metric depth further benefits tissue surface reconstruction, registration, force estimation, and collision avoidance [10], thereby improving intraoperative safety and efficiency.

Despite decades of progress, stereo depth estimation for endoscopy remains challenging. Classical methods [11], [12] are efficient and perform well in controlled settings, yet they struggle with low texture, specular highlights, occlusions, and parameter sensitivity. Learning-based methods, especially stereo matching foundation models, mitigate these issues via 3D cost volumes [13], iterative refinement [14], and hybrid designs [15], which help preserve thin structures and tolerate mild mis-rectification [16]. However, these advances are fragile under the severe domain shifts in surgical scenes, such as non-Lambertian tissue reflectance, scarce texture, and narrow baselines, which undermine correspondence reliability and lead to depth errors [11], [12], [17]. Nevertheless, foundation models carry rich geometric priors from large-scale pretraining, including epipolar consistency and multi-scale smoothness that, when adapted to the surgical domain, regularize correspondence, disambiguate low-texture regions, and improve data efficiency.

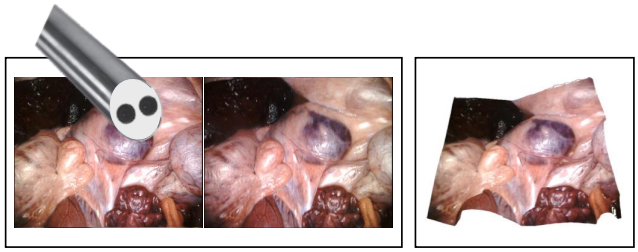

Fig. 1. (a) Laparoscopic images from a dual-channel endoscope. The probe sketch shows two optical paths. (b) A 3D surface reconstructed from stereo matching depth, which can be used for downstream tasks, such as the pre-to-intraoperative registration [18]–[20].

A promising direction is to start from a foundation stereo backbone trained on large, diverse data and adapt it to surgical domains via parameter-efficient fine-tuning. Low-Rank Adaptation (LoRA) [21] is appealing because it freezes the backbone and learns a small set of low-rank parameters, enabling data- and compute-efficient adaptation in clinical pipelines [22]. However, standard LoRA uses a fixed global scaling that is agnostic to camera intrinsics and extrinsics. In stereo endoscopy, the effective disparity scale is governed by the imaging geometry and can drift intraoperatively due to thermal and mechanical effects [23], [24]. These effects motivate a scale-aware adaptation mechanism and a more robust supervision signal that goes beyond simple disparity matching.

In this paper, we propose a camera-aware adaptation framework within geometric constraint for endoscopic stereo depth estimation on a foundation backbone [25]. More specifically, we introduce Camera-Aware LoRA (CaLoRA), which derives a camera-specific gate from focal length and

*This work was supported in part by the National Natural Science Fund for Excellent Young Scientists Fund Program (Overseas) under Grant 22IAA01849. This work was also supported by the National Natural Science Foundation of China under Grant 62303275, Jinan Science and Technology Bureau under Grant 202333011, and the Postdoctoral Fellowship Program of China Postdoctoral Science Foundation under Grant GZC20251184. (†Corresponding author: Zhe Min. Email: minzhe@sdu.edu.cn).

[1]Shixing Ma, Shuwei Shao, Xinzhe Du, Rui Song, Yibin Li, and Zhe Min are with the School of Control Science and Engineering, Shandong University, Jinan 250061, China.

[2]Zhaoxin Lin is with the School of Mechanical Engineering, Tianjin University, Tianjin 300354, China.

[3]Max Q.-H. Meng is with Shenzhen Key Laboratory of Robotics Perception and Intelligence and the Department of Electronic and Electrical Engineering at Southern University of Science and Technology in Shenzhen, China.

[4]Zhe Min is also with the UCL Hawkes Institute and Department of Medical Physics and Biomedical Engineering, University College London, London WC1E 6BT, UK.

baseline and injects it into linear and grouped convolutions. Meanwhile, we develop a dual-view geometric supervision that elevates left-view disparity to 3D and reprojects it to the right view to enforce depth consistency on each view's native grid with a robust Charbonnier loss. We further add a dynamic spectral alignment loss based on a compact learnable Fourier bank to refine tissue boundaries. In order to fully exploit the strengths of these components, we integrate CaLoRA with the geometric and spectral terms into a single training objective for robust endoscopic stereo depth estimation. Fig. 2 illustrates the full pipeline.

To summarize, the contributions of this paper are listed as follows:

- We propose a Camera-Aware LoRA for fine-tuning stereo matching foundation models on surgical data, which normalizes the update magnitude of delta weight using a camera-specific factor derived from the current frame, stabilizing adaptation under intraoperative micro-drift of intrinsics.
- We propose a dual-view geometric constraint that enforces stereo consistency by supervising depth on both views in their native grids via 3D lifting and cross-view reprojection.
- We introduce a lightweight spectral alignment regularizer in loss founction, built on a compact, learnable Fourier bank, that reduces depth ambiguities, preserves thin structures and sharp tissue boundaries, and yields an accurate intraoperative point cloud.

II. Related Work

In this section, we briefly review traditional stereo matching methods, learing-based stereo matching methods and LoRAs for surgical foundation model adaption.

Traditional Stereo Matching. Classical stereo matching was structured into a four-stage pipeline of matching-cost computation, cost aggregation, disparity optimization, and disparity refinement, which established a taxonomy and benchmark for comparative evaluation [26]. At the cost level, non-parametric transforms such as Rank and Census [27] leveraged local intensity orderings to improve robustness to radiometric changes. For explicit occlusion reasoning within the optimization stage, dynamic programming on the disparity-space image (DSI) [28] jointly selected matches and occlusions along each scanline, often stabilized by control points. Moving beyond purely local optimization, Semi-Global Matching (SGM) [11] combined pixel-wise Mutual Information with multi-directional path aggregation and included subpixel refinement and occlusion handling, enabling scalability to larger images at moderate runtime. To accelerate high resolution stereo methods, ELAS [12] introduced sparse support points and Delaunay triangulation to impose a piecewise-linear prior that guided local matching. In contrast, graph-cut methods [29] optimized a global energy to improve matching accuracy in low-texture areas, at the cost of higher computation and memory.

Learning-based Stereo Matching. Recent stereo methods developed from cost volume filtering to iterative refinement [30], and further to hybrid designs that incorporated monocular or foundation priors. Early cost volume-based approaches such as PSMNet [13] constructed 3D CNNs to aggregate contextual information, but the memory and computation requirements grew rapidly at high resolution. Iterative refinement methods such as RAFT-Stereo [14] replaced explicit 3D filtering with correlation pyramids and recurrent updates, which reduced computational demand and improved generalization across datasets, though the iterative nature introduced latency and limited long-range reasoning. Hybrid designs such as CREStereo [15] and Sea-Raft [31] combined cost filtering with recurrence through adaptive or deformable correlation, captured fine structures, and handled mild misrectification at the expense of increased complexity. Semi-supervised designs such as BiSS-DBCNN [32] introduced dual-branch consistency and confidence-aware supervision to exploit unlabeled endoscopic data. More recent works integrated monocular and foundation priors, as in MonSter [33], DEFOM-Stereo [34], and FoundationStereo [25], which strengthened zero-shot generalization by leveraging monocular cues and large-scale training data. Nevertheless, deployment in medical endoscopy remains challenging due to domain shifts introduced by specular highlights, non-Lambertian tissue reflectance, limited texture, and narrow stereo baselines [32], [35], [36]. Unlike prior applications in natural scenes, we adapt stereo matching foundation models to endoscopy in this work.

LoRA for Surgical Model Adaptation. LoRA emerged as an efficient strategy to adapt large-scale foundation models by freezing core weights and introducing low-rank parameter updates [21]. Recent work extended this idea to medical and vision tasks. EndoARSS [37] integrated LoRA into DINOv2 [38] transformers for multitask endoscopic surgery analysis, combining activity recognition and semantic segmentation under a unified framework. EndoDAC [22] introduced Dynamic Vector-based LoRA (DV-LoRA) to adapt monocular depth foundation models for self-supervised endoscopic depth estimation, while Endo3DAC [39] further proposed Gated Dynamic Vector-based LoRA (GDV-LoRA) for efficient joint depth, pose, and camera intrinsics estimation from surgical videos. In surgical scene segmentation, LoRASAM [40] adapted the Segment Anything Model (SAM) using LoRA layers and text-based prompts to reduce trainable parameters while addressing domain shift from natural to surgical images. Surgical-LVLM [41] incorporated LoRA to adapt large vision-language models for grounded surgical question answering, demonstrating its flexibility in multimodal reasoning. Collectively, these works illustrated how LoRA and its variants adapted vision, depth, segmentation, and multimodal models for surgical and general visual domains with reduced computational cost. However, previous work has paid limited attention to the stereo matching task, where correspondences are constrained by epipolar geometry. To this end, we propose CaLoRA to adapt the geometric priors in stereo matching foundation models to the stereo endoscopic domain.

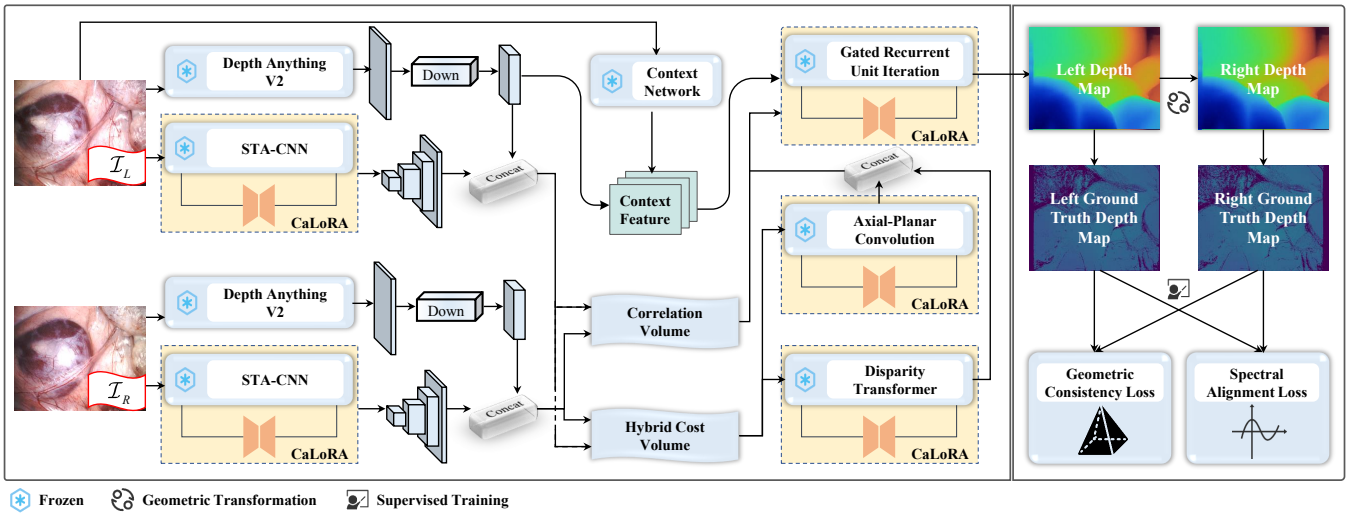

Fig. 2. Framework of the proposed method. The frozen DepthAnythingV2 provides monocular priors via a Side-Tuning Adapter (STA), which extracts the input image context. The Axial-Planar Convolution (APC) module filters the 4D cost volume by decoupling spatial and disparity convolutions. The Disparity Transformer (DT) enhances long-range context within the cost volume using multihead self-attention, its output is upsampled and added element-wise to the APC output. An iterative Convolutional Gate Recurrent Unit (ConvGRU) performs refinement of the disparity map to produce the final prediction.

III. Methodology

In this section, we first present CaLoRA for scale-adaptive tuning and its injection method into linear and convolutional layers of the stereo backbone. Then, we introduce dual-view geometric supervision to enforce the depth consistency on each view's native grid. Finally, we propose a dynamic spectral alignment loss and integrate it with the geometric term into a single objective for robust endoscopic stereo depth estimation.

A. Camera-Aware LoRA for Scale-Adaptive Model Tuning

The CaLoRA Formulation. For rectified pairs, the scene scale Z is governed by focal length f and baseline b, via $Z \propto \frac{fb}{d}$. We construct a camera-aware scalar gate γ from per-sample camera scales and inject it into LoRA. For each sample i in a batch, we define the scale ratio:

$$r_i = \frac{f_i b_i}{s_0}, \tag{1}$$

where $s_0 = \mathrm{median}(\{f_j\}) \cdot \mathrm{median}(\{b_j\})$ is a robust batch reference, and $f_i = (K_1)_{11}$ and b_i denote the effective focal length and baseline, respectively. This ratio r_i is then mapped to a preliminary gate value ψ_i:

$$\psi_i = \sigma\big(\beta \log r_i\big) = \frac{r_i^{\beta}}{1 + r_i^{\beta}} \in (0, 1), \tag{2}$$

where $\sigma(x) = 1/(1+e^{-x})$ is the sigmoid function, and $\beta \in \mathbb{R}_{>0}$. The Eq. (2) has following properties: it is monotone (for any $r_1, r_2 > 0$ with $r_1 < r_2$, $\psi(r_1) < \psi(r_2)$), symmetric around the reference scale ($\psi(1/r) = 1 - \psi(r)$), and has a bounded gradient ($\frac{\partial \psi}{\partial r} = \frac{\beta}{r}\psi(1-\psi) \le \frac{\beta}{4r}$), which ensures stable optimization. Finally, we aggregate these per-sample factors into a single batch scalar and broadcast it to all LoRA modules:

$$\gamma = \frac{1}{B}\sum_{i=1}^{B} \psi_i. \tag{3}$$

Injection into CaLoRA. Given a frozen core weight W_{core} and a rank-r update $\Delta W = BA$, CaLoRA scales the low-rank injection by this camera-aware gate γ:

$$W_{\text{eff}} = W_{\text{core}} + \frac{\alpha}{r}\,\gamma\, BA, \tag{4a}$$

$$y = x\, W_{\text{core}}^{\top} \; + \; \frac{\alpha}{r}\,\gamma\, xA^{\top}B^{\top}. \tag{4b}$$

Properties of CaLoRA. Since $0 < \gamma < 1$, the injected update is uniformly bounded:

$$\big\|W_{\text{eff}} - W_{\text{core}}\big\|_F = \frac{\alpha}{r}\,\gamma\,\|BA\|_F \;\le\; \frac{\alpha}{r}\,\|BA\|_F. \tag{5}$$

The bounded gradient of the mapping function ψ ensures a smooth dependence on camera scale. While the overall update magnitude of delta weight is governed by the LoRA hyperparameter α, the adaptive gate γ is responsible for shaping the relative strength of the update as the camera scale varies.

Convolutional Operators in CaLoRA. CaLoRA applies to any linear operator whose kernel can be flattened into a matrix. This generality is essential for stereo backbones, which are convolution-heavy and rely on 2D and 3D Convolutional layers, including grouped and depthwise variants. For a grouped Conv$n$D layer with G groups, let c_{in} and c_{out} denote the input and output channel counts, respectively. The spatial kernel contains $k_{\text{sp}} = \prod_d k_d$ elements ($k_H k_W$ for 2D and $k_D k_H k_W$ for 3D). The per-group flattened core weight is defined as:

$$W_{\text{core}}^{(g)} \in \mathbb{R}^{\frac{c_{\text{out}}}{G} \times \left(\frac{c_{\text{in}}}{G}\, k_{\text{sp}}\right)}.$$

CaLoRA then introduces a low-rank update $\Delta W^{(g)} = B^{(g)}A^{(g)}$ with $A^{(g)} \in \mathbb{R}^{r \times \left(\frac{c_{\text{in}}}{G} k_{\text{sp}}\right)}$ and $B^{(g)} \in \mathbb{R}^{\frac{c_{\text{out}}}{G} \times r}$, which is reshaped back to the original kernel layout at inference time. Finally, CaLoRA yields the effective weight:

$$W_{\text{eff}}^{(g)} = W_{\text{core}}^{(g)} + \frac{\alpha}{r}\,\gamma\,\Delta W^{(g)}, \tag{6}$$

which is reshaped back to the native tensor shape and concatenated across the G groups along channels. The batch-level scalar γ is broadcast uniformly to all LoRA modules, consistently modulating the strength of low-rank updates across 2D/3D convolutions.

Architectural Integration with CaLoRA. To preserve global reasoning and monocular priors, we do not apply LoRA to the Vision Transformer (ViT) branch. As shown in Fig. 2, we apply CaLoRA at four sites, from high-level priors to low-level iterative refinement:

1) The STA-CNN layers at $1/16$ and $1/32$ scales to softly adapt fused geometric priors while keeping the ViT backbone in DepthAnythingV2 [42] branch frozen.
2) The APC hourglass 3D convolutions that decouples spatial and disparity processing ($K_s{\times}K_s{\times}1$; $1{\times}1{\times}K_d$) to modulate the spatial disparity aggregation trade-off and enlarge the effective disparity receptive field with negligible overhead.
3) The $4{\times}4{\times}4$ stride-4 3D convolutions that downsamples the cost volume before tokenization and transformer encoding to calibrate the statistics and scale of the input to Disparity Transformer (DT) under camera-scale variations, while keeping the DT encoder frozen to preserve long-range reasoning.
4) The ConvGRU-based iterative refinement stack, including gated convolutions and the regression head, to directly modulate per-step update magnitude and direction.

B. Dual-View Geometry with Spectral Structural Regularization

We consider a calibrated stereo pair $(\mathcal{I}_L, \mathcal{I}_R)$ with intrinsics $K_1, K_2 \in \mathbb{R}^{3\times3}$ and the right-camera pose (R, t) w.r.t. the left. The left-view pixel domain, comprising all image coordinates (u, v), is denoted by $\Omega_L = \{(u, v) \mid 0 \le u \le W, 0 \le v \le H\}$. From this, the network predicts a left-view disparity map $\hat{d} : \Omega_L \in \mathbb{R}_{>0}$, and the right camera provides different visibility with more geometric feature. Our supervision leverages two complementary terms: a geometric loss that compares depth in each view's own domain, and a spectral loss that matches structural fingerprints across scales.

Geometric Consistency via Left-to-Right Projection. Given the left-view disparity $\hat{d}$, a rectified pair with effective focal length f_x and baseline B, the predicted left-view metric depth is directly obtained from the disparity: $\hat{D}_L(u, v) = \frac{f_x B}{\hat{d}(u,v)}$. Next, we warp this prediction to the right view through a chain of geometric transformations. For each left-view pixel (u, v), we perform the following steps:

1) The left-view depth $\hat{D}_L$ is used to back-project the pixel (u, v) into a 3D point $\mathbf{p}_L$ in the left camera's frame, and $\mathbf{p}_L(u, v) = \hat{D}_L(u, v)\, K_1^{-1}[u,\, v,\, 1]^\top$.
2) The 3D point is transformed into the right camera's coordinate system using the known pose (R, t), and $\mathbf{p}_R(u, v) = R\,\mathbf{p}_L(u, v) + t$.
3) The transformed 3D point $\mathbf{p}_R$ is projected onto the right image plane using its intrinsics K_2 to obtain the corresponding, potentially non-integer coordinates, and $(u', v') = \Pi(K_2\mathbf{p}_R(u, v))$, where $\Pi([x, y, z]^\top) = (x/z, y/z)$.

In this stage, the z-component of $\mathbf{p}_R$, denoted as $(\mathbf{p}_R)_z$, represents the precise metric depth associated with the non-integer coordinate (u', v'). To obtain the final predicted right-view depth map, $\hat{D}_R$, which is defined on a regular pixel grid, we use differentiable bilinear sampling to retrieve values at the integer grid locations based on these projected points. With both $\hat{D}_L$ and $\hat{D}_R$ established, we define geometric loss using the Charbonnier penalty $\rho_\varepsilon(r) = \sqrt{r^2 + \varepsilon^2}$, with $\varepsilon > 0$ [43], [44]. The loss is averaged over whole left-view pixels (Ω_L) and all valid right-view projections. We define the set of source pixels for these valid right-view projections as Ω_R, which includes all pixels (u, v) from the left-view domain, and the warped coordinates (u', v') land within the image bounds, while $\Omega_R = \{(u, v) \in \Omega_L \mid 0 \le u' < W, 0 \le v' < H, (\mathbf{p}_R(u, v))_z > 0\}$. The geometric loss is formulated as:

$$\mathcal{L}_{\text{base}} = \left\langle \rho_\varepsilon\big(\hat{D}_L - D_L^{\text{gt}}\big) \right\rangle_{\Omega_L} + \left\langle \rho_\varepsilon\big(\hat{D}_R - D_R^{\text{gt}}\big) \right\rangle_{\Omega_R}, \quad (7)$$

where ground-truth depths D_L^{gt} and D_R^{gt} are obtained from the calibrated ground truth point cloud, and $\langle\cdot\rangle_\Omega$ denotes the arithmetic average of the enclosed term over the index set. The second term in $\mathcal{L}_{\text{base}}$ is critical, it is not a mere change of variables but supervises the prediction along a different set of ray directions on a different grid, leveraging the unique visibility of the right camera as independent geometric information.

Structural Fidelity via Dynamic Spectral Alignment. Inspired by prior work on spectral losses [45], [46], we introduce a complementary loss $\mathcal{L}_{\text{bank}}$, which moves beyond individual pixel values to match the multi-scale structural fingerprint of the ground truth. The process begins by normalizing each depth value x to a canonical phase range $[-\pi, \pi]$ to ensure the mapping is comparable across datasets:

$$\tilde{x} = \pi\left(2\,\frac{x - D_{\min}}{D_{\max} - D_{\min}} - 1\right). \quad (8)$$

This normalized phase $\tilde{x}$ is then transformed into a $2K$-dimensional feature vector $\Phi_K(x)$ using a bank of sin and cos functions at learnable frequencies α_k (where $f_k = \pi\, 2^{\alpha_k}$), allowing the model to adaptively focus on the most informative scales:

$$\Phi_K(x) = \big[\sin(f_k\tilde{x}),\, \cos(f_k\tilde{x})\big]_{k=1}^{K} \in \mathbb{R}^{2K}. \quad (9)$$

To ensure stability and prioritize large-scale structures, each component of this feature vector is weighted by a dynamic regularizer, $w_k = \frac{1}{1+f_k^2}$. We use this to define a general function, $E(D^p, D^g; \Theta)$, which computes the robust spectral distance between a predicted depth map D^p and a ground-truth map D^g over a given pixel domain Θ:

$$E(D^p, D^g; \Theta) = \left\langle \rho_\varepsilon \| W^{1/2}\big(\Phi_K(D^p) - \Phi_K(D^g)\big) \|_2 \right\rangle_\Theta, \quad (10)$$

where W is the diagonal matrix of weights w_k. Finally, similar to geometric loss in Eq. (7), Eq. (10) is applied to

TABLE I

QUANTITATIVE RESULTS ON THE SCARED AND HAMLYN DATASETS. LOWER IS BETTER FOR RMSE, RMSE LOG, EPE, ABS REL, AND D1 (↓).

Method	SCARED					Hamlyn					Time (s)↓
	RMSE↓	RMSE log↓	EPE↓	Abs Rel↓	D1↓	RMSE↓	RMSE log↓	EPE↓	Abs Rel↓	D1↓	
Raft-Stereo [14]	5.491	0.078	4.507	0.065	0.430	4.579	0.067	3.005	0.047	0.310	0.108
Sea-Raft [31]	6.006	0.087	4.649	0.069	0.438	4.593	0.067	2.922	0.046	0.284	**0.056**
CREStereo [15]	5.458	0.078	4.514	0.065	0.437	4.422	0.065	2.902	0.046	0.295	0.216
PSMNet [13]	5.479	0.079	4.394	0.063	0.430	6.424	0.094	3.812	0.057	0.339	0.129
Defom-Stereo [34]	5.832	0.082	4.885	0.070	0.458	4.842	0.072	3.031	0.049	0.302	0.203
BiSS-DBCNN [32]	5.439	0.078	4.478	0.065	0.428	4.321	0.063	2.787	0.044	0.270	0.097
MonSter [33]	4.957	0.070	4.113	0.060	0.380	8.640	0.106	4.068	0.054	0.281	6.344
FoundationStereo [25]	5.394	0.077	4.405	0.064	0.423	4.655	0.067	2.939	0.046	0.284	1.140
Ours	**3.830**	**0.054**	**2.731**	**0.039**	**0.234**	**4.215**	**0.055**	**2.301**	**0.036**	**0.177**	0.218

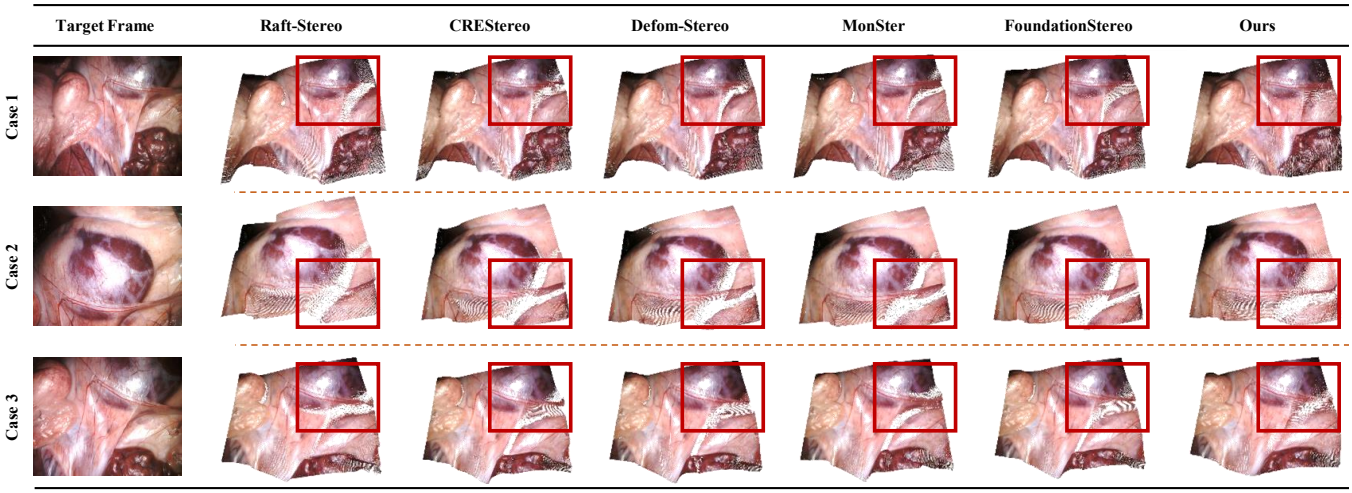

Fig. 3. Qualitative comparison of point clouds reconstructed from disparity. Columns show reconstructions from the same input frame using different methods. CaLoRA-Stereo reconstructs point set that are smooth and continuous while preserving boundary details. Red boxes highlight defects in the predicted results.

both the left and right views:

$$\mathcal{L}_{\text{bank}} = E(\hat{D}_L, D_L^{\text{gt}}; \Omega_L) + E(\hat{D}_R, D_R^{\text{gt}}; \Omega_R). \tag{11}$$

Total Optimizing Objective. After warping depth map to the right-view grid and comparing them in the spectral domain, we enforce both geometric and structural consistencies on each view's native grid. Consequently, the total training objective is a weighted sum of the geometric and spectral losses:

$$\mathcal{L} = w_{\text{base}}\, \mathcal{L}_{\text{base}} + w_{\text{bank}}\, \mathcal{L}_{\text{bank}}, \tag{12}$$

where w_{base} and w_{bank} are weighting coefficients that balance the geometric consistency loss and the spectral alignment loss.

IV. EXPERIMENTS

A. Implementation Details

Datasets. We use two medical datasets of laparoscopic surgery, and the depth map is capped at 150mm. The SCARED benchmark [47], introduced in the MICCAI 2019 challenge, consists of 35 laparoscopic stereo video sequences (27,826 frames) of freshly dissected porcine abdominal specimens, acquired with a *da Vinci Xi* surgical endoscope. For both pre-training and evaluation, we follow the default dataset split, and resize all video frames to a resolution of 320×256 pixels. The Hamlyn dataset [48], [49] consists of laparoscopic and endoscopic video sequences from a variety of surgical procedures, as well as a heart phantom sequence with associated point cloud ground truth. It provides challenging in vivo scenarios characterized by frequent occlusions and specular highlights. We report results on a validation subset of heart videos.

Training Details. All experiments are implemented in PyTorch and executed on a workstation with Intel Xeon Gold 6154 CPUs and NVIDIA RTX 4090 GPUs. We train for 200 epochs using AdamW (learning rate 5×10^{-6}) with a total batch size of 8. The ConvGRU update module is unrolled for 12 iterations per forward pass. CaLoRA uses a scaling factor $\alpha = 16$ with $r = 8$ and $\beta = 1$. The camera-aware scalar γ is learned and constrained to $[0, 1]$ by sigmoid founction. In the total optimization objective, we set $w_{\text{base}} = 0.6$ and $w_{\text{bank}} = 0.1$. For the Charbonnier penalty in Eq. (7), we set $\varepsilon = 10^{-3}$. In the Fourier bank, we use $K = 8$ frequencies with learnable $\alpha_k \in [\alpha_{\min}, \alpha_{\max}] = [-3, 3]$, and a depth normalization range of $D_{\min} = 0.1$ and $D_{\max} = 3.0$.

Evaluation Metrics. We evaluate all methods using five standard depth estimation metrics: RMSE, RMSE log, EPE, Abs Rel, and D1 [50]–[52]In addition, we report the average inference time per frame (Time) in second, measured during the evaluation stage. The root mean square error (RMSE) is

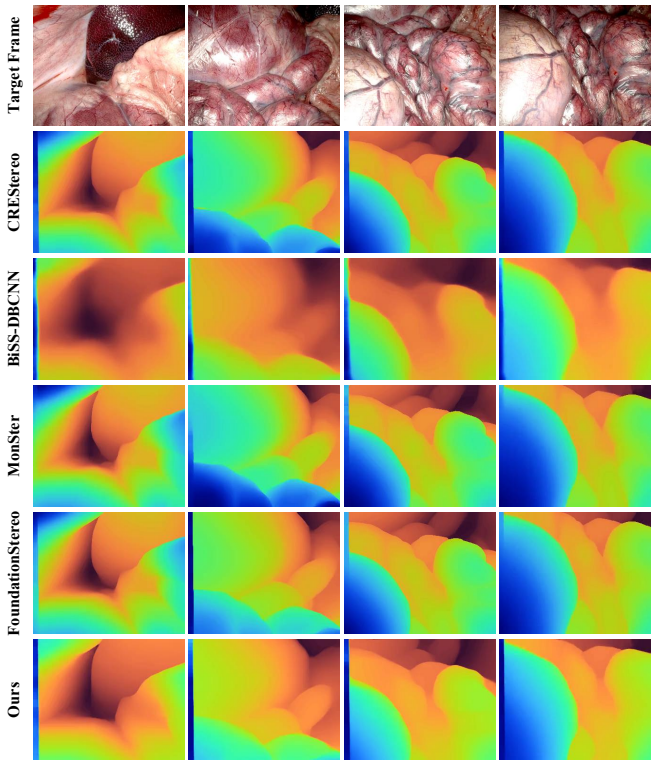

Fig. 4. Qualitative comparison of pseudocolor disparity maps on the SCARED dataset. After stereo rectification, a margin appears on the left edge of the left image, those pixels are invalid.

defined as

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(d_i^{\text{pred}} - d_i^{\text{gt}}\right)^2}, \tag{13}$$

which measures the average Euclidean error between predicted and ground truth depths. The logarithmic root mean square error (RMSE log) is defined as

$$\text{RMSE}_{\log} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\log d_i^{\text{pred}} - \log d_i^{\text{gt}}\right)^2}, \tag{14}$$

which compares depths in log space and emphasizes relative errors. The end-point error (EPE) is given by

$$\text{EPE} = \frac{1}{N}\sum_{i=1}^{N}|d_i^{\text{pred}} - d_i^{\text{gt}}|, \tag{15}$$

which corresponds to the mean absolute deviation between prediction and ground truth. The absolute relative error (Abs Rel) is defined as

$$\text{AbsRel} = \frac{1}{N}\sum_{i=1}^{N}\frac{|d_i^{\text{pred}} - d_i^{\text{gt}}|}{d_i^{\text{gt}}}, \tag{16}$$

which captures the average relative deviation of predictions with respect to the ground truth. The D1 metric quantifies the fraction of points whose errors simultaneously exceed both an absolute threshold of 3 mm and a relative threshold of 5%, to reflect the proportion of mispredicted pixels.

B. Experimental Results Analysis

Table I reports experimental results on SCARED and Hamlyn against representative stereo matching methods and prior-enhanced models. We select the second best method to compare the quantitative difference. All pipelines were reproduced under the same experimental conditions and evaluated with identical data splits, preprocessing, and metrics to ensure a fair comparison.

SCARED. Our method surpasses previous competing approaches on most evaluated metrics except reasoning Time. Taking MonSter as a reference, we compare CaLoRA-Stereo on each metric. RMSE decreases by 22.7% from 4.957 to 3.830. RMSE log decreases by 22.9% from 0.070 to 0.054. EPE decreases by 33.6% from 4.113 to 2.731. Abs Rel decreases by 35.0% from 0.060 to 0.039. D1 decreases by 38.4% from 0.380 to 0.234. The macro-average relative reduction over the five metrics is 30.5%. These improvements are statistically significant according to a two-sided paired t-test ($p < 0.001$). For visualization comparison, as shown in Fig. 3, CaLoRA-Stereo reconstructs an intact tissue surface whereas the other method miss fine details in Case 1. In Case 2, the compared methods exhibit clear faults while our prediction remains continuous, and CaLoRA-Stereo produces the fewest holes in Case 3. As shown in Fig. 4, disparity maps further highlight our approach's performance in detail recovery and depth consistency.

Hamlyn. Our method also exceeds previous leading approaches on the Hamlyn dataset. Compared with BiSS-DBCNN, our RMSE decreases by 2.5% from 4.321 to 4.215. RMSE log decreases by 12.7% from 0.063 to 0.055. EPE decreases by 17.4% from 2.787 to 2.301. Abs Rel decreases by 18.2% from 0.044 to 0.036. D1 decreases by 34.4% from 0.270 to 0.177. The mean relative reduction is 17.0%. The improvements are most pronounced on outlier-sensitive measures, namely D1, EPE, and Abs Rel. Significance holds under a two-sided paired t-test ($p < 0.001$). As illustrated in Fig. 5, the qualitative results indicate that CaLoRA-Stereo produces fewer outliers and fewer depth estimation failures.

Efficiency. CaLoRA-Stereo runs at 0.218s per frame, on par with CREStereo (0.216s). Crucially, it is more efficient than other stereo foundation models, running over 5.2x faster than the original FoundationStereo and 29x faster than MonSter.

C. Ablation Studies

To better understand the effect of each key component in our framework, we conduct an ablation study and present the results in Table II.

We propose a CaLoRA, which is improved upon standard LoRA, with a camera-aware scaling mechanism. Under identical training and evaluation settings (r=8) in the ablation study, CaLoRA consistently outperforms standard LoRA across all metrics, confirming the effectiveness of the proposed modification and its scale consistency under small intrinsics drift. While we examine CaLoRA's rank sensitivity with $\mathcal{L}_{\text{base}}$ and $\mathcal{L}_{\text{bank}}$, it is suboptimal at r=1, peaks at r=8, degrades at r=16, and collapses at r=32. We

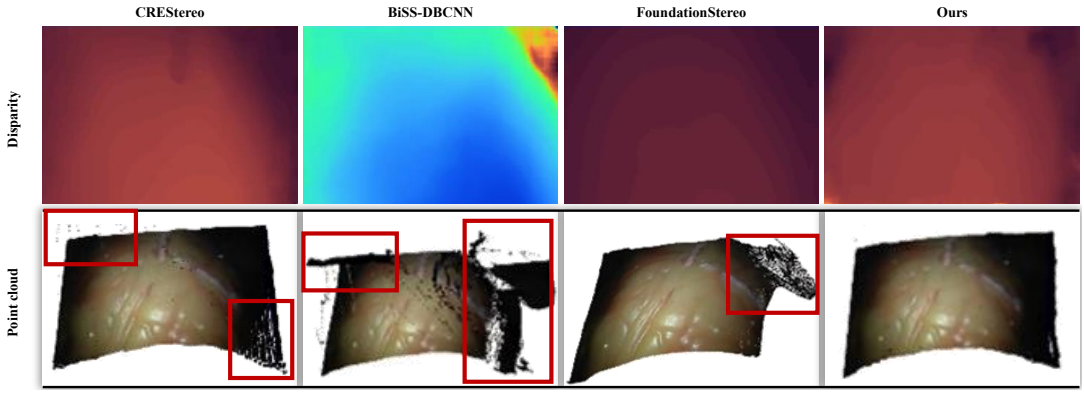

Fig. 5. Qualitative comparison of pseudocolor disparity maps on the Hamlyn dataset. Red boxes highlight outliers and severe failure areas.

TABLE II

COMPONENT AND RANK ABLATIONS. LOWER IS BETTER FOR ALL METRICS (↓).

LoRA	CaLoRA	$\mathcal{L}_{\text{base}}$	$\mathcal{L}_{\text{bank}}$	r	RMSE↓	RMSE log↓	EPE↓	Abs Rel↓	D1↓
Component ablation (fixed LoRA rank r=8)									
✓		✓	✓	8	3.986	0.056	3.247	0.046	0.313
	✓	✓		8	3.547	0.051	2.778	0.041	0.262
	✓	✓	✓	8	**3.067**	**0.043**	**2.328**	**0.033**	**0.206**
Ablation on CaLoRA rank r									
	✓	✓	✓	1	3.345	0.048	2.572	0.038	0.237
	✓	✓	✓	4	3.626	0.051	2.855	0.041	0.272
	✓	✓	✓	16	3.801	0.054	3.042	0.044	0.292
	✓	✓	✓	32	8.846	0.126	7.858	0.108	0.654

therefore adopt r=8 as a principled trade-off option. We then remove the spectral term $\mathcal{L}_{\text{bank}}$ from the full configuration and observe a significant degradation in depth estimation accuracy, confirming that multi-scale spectral alignment improves accuracy by enforcing structural consistency.

V. CONCLUSIONS

In this work, we present a parameter-efficient fine-tuning framework for stereo matching foundation model that leverages inherent geometric priors to improve depth estimation accuracy in endoscopic scenes. In parallel, we propose a dual-view geometric supervision lifts the left-view prediction to 3D and reprojects it to the right, allowing each view to be supervised on its native grid with its own visibility. Besides, we introduce a lightweight spectral alignment regularizer based on a small learnable Fourier bank that further preserves fine tissue surface details. The method is plug-and-play for linear and convolutional layers, and requires no backbone modification. Experiments on SCARED and Hamlyn indicate uniformly lower errors across standard metrics with practical runtime. Looking ahead, the same mechanism has potential to support cross-device and cross-hospital deployment, lens or coupler swaps with different magnifications or view angles, optical or digital zoom and cropping in the imaging chain, and intraoperative scope replacement or reinsertion.

REFERENCES

[1] Y. Long, A. Lin, D. H. C. Kwok, L. Zhang, Z. Yang, K. Shi, L. Song, J. Fu, H. Lin, W. Wei *et al.*, "Surgical embodied intelligence for generalized task autonomy in laparoscopic robot-assisted surgery," *Science Robotics*, vol. 10, no. 104, p. eadt3093, 2025.

[2] Z. Min, J. Lai, and H. Ren, "Innovating robot-assisted surgery through large vision models," *Nature Reviews Electrical Engineering*, vol. 2, no. 5, pp. 350–363, 2025.

[3] H. Saeidi, J. D. Opfermann, M. Kam, S. Wei, S. Léonard, M. H. Hsieh, J. U. Kang, and A. Krieger, "Autonomous robotic laparoscopic surgery for intestinal anastomosis," *Science robotics*, vol. 7, no. 62, p. eabj2908, 2022.

[4] Y. Ma, X. An, Q. Yang, M. Cai, Z. Tang, J. Chang, V. Iacovacci, T. Xu, L. Zhang, and Q. Wang, "Magnetic continuum robot for intelligent manipulation in medical applications," *SmartBot*, vol. 1, no. 2, p. e12011, 2025.

[5] Y. Li, S. Ma, Z. Yang, S. Jiang, Z. Lin, and Z. Zhou, "A multi-optical and mechanical compensation robotic surgery system based on augmented reality for endoscopic neurosurgery," *Journal of Medical Devices*, vol. 19, no. 2, p. 021005, 12 2024. [Online]. Available: https://doi.org/10.1115/1.4067172

[6] Z. Zhou, Z. Yang, S. Jiang, J. Zhuo, Y. Li, T. Zhu, S. Ma, and J. Zhang, "Validation of a surgical navigation system for hypertensive intracerebral hemorrhage based on mixed reality using an automatic registration method," *Virtual Reality*, vol. 27, no. 3, pp. 2059–2071, 2023.

[7] S. Ma, Q. Wang, X. Du, A. Zhang, Q. Jin, Y. Liu, Q. Yin, W. Liu, R. Song, Y. Li *et al.*, "A comparative study of augmented reality-assisted orthopedic surgical navigation systems," in *2025 IEEE International Conference on Real-time Computing and Robotics (RCAR)*. IEEE, 2025, pp. 959–964.

[8] A. Perez, H. Zhang, Y.-C. Ku, L. Seenivasan, R. Soberanis, J. L. Porras, R. Day, J. Jopling, P. Najjar, and M. Unberath, "Privacy-preserving operating room workflow analysis using digital twins," *arXiv preprint arXiv:2504.12552*, 2025.

[9] S. Shao, Z. Pei, W. Chen, P. C. Chen, and Z. Li, "Nddepth: Normal-distance assisted monocular depth estimation and completion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 8883–8899, 2024.

[10] A. Shademan, R. S. Decker, J. D. Opfermann, S. Leonard, A. Krieger, and P. C. Kim, "Supervised autonomous robotic soft tissue surgery," *Science translational medicine*, vol. 8, no. 337, pp. 337ra64–337ra64, 2016.

[11] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 328–341, 2007.

[12] A. Geiger, M. Roser, and R. Urtasun, "Efficient large-scale stereo matching," in *Asian conference on computer vision*. Springer, 2010, pp. 25–38.

[13] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5410–5418.

[14] L. Lipson, Z. Teed, and J. Deng, "Raft-stereo: Multilevel recurrent field transforms for stereo matching," in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021, pp. 218–227.

[15] J. Li, P. Wang, P. Xiong, T. Cai, Z. Yan, L. Yang, J. Liu, H. Fan, and S. Liu, "Practical stereo matching via cascaded recurrent network with adaptive correlation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 263–16 272.

[16] L. Bai, T. Chen, Q. Tan, W. J. Nah, Y. Li, Z. He, S. Yuan, Z. Chen, J. Wu, M. Islam *et al.*, "Endouic: Promptable diffusion transformer for unified illumination correction in capsule endoscopy," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 296–306.

[17] X. Song, X. Zhao, L. Fang, H. Hu, and Y. Yu, "Edgestereo: An effective multi-task learning network for stereo matching and edge detection," *International Journal of Computer Vision*, vol. 128, no. 4, pp. 910–930, 2020.

[18] X. Du, S. Ma, Z. Zhang, R. Song, Y. Li, M. Q.-H. Meng, and Z. Min, "Registration after completion: Towards sparse and partial point set registration for computer-assisted orthopedic surgery," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025, pp. 7710–7717.

[19] Z. Min, Z. M. Baum, S. U. Saeed, S. Ma, X. Du, M. Emberton, D. C. Barratt, Z. A. Taylor, and Y. Hu, "Biomechanics-informed non-rigid medical image registration with elasticity theories," *IEEE Transactions on Medical Imaging*, 2026.

[20] Z. Feng, J. Liu, and H. Wang, "Optimizing scene flow with neural rigidity prior," *Robot Learning*, vol. 1, no. 1, pp. 1–15, 2024.

[21] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.

[22] B. Cui, M. Islam, L. Bai, A. Wang, and H. 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*. Springer, 2024, pp. 208–218.

[23] M. Hardner, R. Docea, and D. Schneider, "Guided calibration of medical stereo endoscopes," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 43, pp. 679–686, 2022.

[24] H. Ghasemzadeh and D. D. Deliyski, "Non-linear image distortions in flexible fiberoptic endoscopes and their effects on calibrated horizontal measurements using high-speed videoendoscopy," *Journal of Voice*, vol. 36, no. 6, pp. 755–769, 2022.

[25] B. Wen, M. Trepte, J. Aribido, J. Kautz, O. Gallo, and S. Birchfield, "Foundationstereo: Zero-shot stereo matching," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 5249–5260.

[26] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International journal of computer vision*, vol. 47, no. 1, pp. 7–42, 2002.

[27] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *European conference on computer vision*. Springer, 1994, pp. 151–158.

[28] A. F. Bobick and S. S. Intille, "Large occlusion stereo," *International Journal of Computer Vision*, vol. 33, no. 3, pp. 181–200, 1999.

[29] B. Lu, L. Sun, L. Yu, and X. Dong, "An improved graph cut algorithm in stereo matching," *Displays*, vol. 69, p. 102052, 2021.

[30] Z. Li, X. Liu, N. Drenkow, A. Ding, F. X. Creighton, R. H. Taylor, and M. Unberath, "Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6197–6206.

[31] Y. Wang, L. Lipson, and J. Deng, "Sea-raft: Simple, efficient, accurate raft for optical flow," in *European Conference on Computer Vision*. Springer, 2024, pp. 36–54.

[32] H. Shi, Z. Wang, Y. Zhou, D. Li, X. Yang, and Q. 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*, vol. 42, no. 11, pp. 3269–3282, 2023.

[33] J. Cheng, L. Liu, G. Xu, X. Wang, Z. Zhang, Y. Deng, J. Zang, Y. Chen, Z. Cai, and X. Yang, "Monster: Marry monodepth to stereo unleashes power," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 6273–6282.

[34] H. Jiang, Z. Lou, L. Ding, R. Xu, M. Tan, W. Jiang, and R. Huang, "Defom-stereo: Depth foundation model based stereo matching," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 21 857–21 867.

[35] Z. Yang, R. Simon, and C. A. Linte, "Disparity refinement framework for learning-based stereo matching methods in cross-domain setting for laparoscopic images," *Journal of Medical Imaging*, vol. 10, no. 4, pp. 045 001–045 001, 2023.

[36] Y. Ding, C. Han, S. Du, Y. Wang, and D. Qian, "Lightendostereo: A real-time lightweight stereo matching method for endoscopy images," *arXiv preprint arXiv:2503.00731*, 2025.

[37] G. Wang, R. Tang, M. Xu, L. Bai, H. Gao, and H. Ren, "Endoarss: Adapting spatially aware foundation model for efficient activity recognition and semantic segmentation in endoscopic surgery," *Advanced Intelligent Systems*, p. 2500288, 2025.

[38] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[39] B. Cui, L. Bai, M. Islam, A. Wang, Z. Ma, Y. Huang, F. Li, Z. Chen, Z. Jiang, N. Navab *et al.*, "Learning to efficiently adapt foundation models for self-supervised endoscopic 3d scene reconstruction from any cameras," *arXiv preprint arXiv:2503.15917*, 2025.

[40] J. N. Paranjape, S. Sikder, S. S. Vedula, and V. M. Patel, "Low-rank adaptation of segment anything model for surgical scene segmentation," in *International Conference on Pattern Recognition*. Springer, 2024, pp. 187–202.

[41] G. Wang, L. Bai, W. J. Nah, J. Wang, Z. Zhang, Z. Chen, J. Wu, M. Islam, H. Liu, and H. Ren, "Surgical-lvlm: Learning to adapt large vision-language model for grounded visual question answering in robotic surgery," *arXiv preprint arXiv:2405.10948*, 2024.

[42] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 10 371–10 381.

[43] D. Sun, S. Roth, and M. J. Black, "Secrets of optical flow estimation and their principles," in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 2432–2439.

[44] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 624–632.

[45] J.-H. Lee, M. Heo, K.-R. Kim, and C.-S. Kim, "Single-image depth estimation based on fourier domain analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 330–339.

[46] L. Jiang, B. Dai, W. Wu, and C. C. Loy, "Focal frequency loss for image reconstruction and synthesis," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 13 919–13 929.

[47] M. Allan, J. Mcleod, C. Wang, J. C. Rosenthal, Z. Hu, N. Gard, P. Eisert, K. X. Fu, T. Zeffiro, W. Xia *et al.*, "Stereo correspondence and reconstruction of endoscopic data challenge," *arXiv preprint arXiv:2101.01133*, 2021.

[48] P. Pratt, D. Stoyanov, M. Visentini-Scarzanella, and G.-Z. Yang, "Dynamic guidance for robotic surgery using image-constrained biomechanical models," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2010, pp. 77–85.

[49] D. Stoyanov, M. V. Scarzanella, P. Pratt, and G.-Z. Yang, "Real-time stereo reconstruction in robotically assisted minimally invasive surgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2010, pp. 275–282.

[50] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in neural information processing systems*, vol. 27, 2014.

[51] X. Du, S. Ma, M. Liu, Z. Zhang, and Z. Min, "Point set registration metrics reloaded for computer-assisted surgery," in *International Workshop on Shape in Medical Imaging*. Springer, 2025, pp. 319–334.

[52] S. Shao, Z. Pei, W. Chen, W. Zhu, X. Wu, D. Sun, and B. Zhang, "Self-supervised monocular depth and ego-motion estimation in endoscopy: Appearance flow to the rescue," *Medical image analysis*, vol. 77, p. 102338, 2022.