IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

StereoMamba: Real-Time and Robust Intraoperative Stereo Disparity Estimation via Long-Range Spatial Dependencies

Xu Wang, Jialang Xu, *Graduate Student Member, IEEE*, Shuai Zhang, Baoru Huang, *Member, IEEE*, Danail Stoyanov, *Fellow, IEEE*, and Evangelos B. Mazomenos, *Member, IEEE*

***Abstract*—Stereo disparity estimation is crucial for obtaining depth information in robot-assisted minimally invasive surgery (RAMIS). While current deep learning methods have made significant advancements, challenges remain in achieving an optimal balance between accuracy, robustness, and inference speed. To address these challenges, we propose the StereoMamba architecture, which is specifically designed for stereo disparity estimation in RAMIS. Our approach is based on a novel Feature Extraction Mamba (FE-Mamba) module, which enhances long-range spatial dependencies both within and across stereo images. To effectively integrate multi-scale features from FE-Mamba, we then introduce a novel Multidimensional Feature Fusion (MFF) module. Experiments against the state-of-the-art on the ex-vivo SCARED benchmark demonstrate that StereoMamba achieves superior performance on EPE of 2.64 px and depth MAE of 2.55 mm, the second-best performance on Bad2 of 41.49% and Bad3 of 26.99%, while maintaining an inference speed of 21.28 FPS for a pair of high-resolution images (1280 × 1024), striking the optimum balance between accuracy, robustness, and efficiency. Furthermore, by comparing synthesized right images, generated from warping left images using the generated disparity maps, with the actual right image, StereoMamba achieves the best average SSIM (0.8970) and PSNR (16.0761), exhibiting strong zero-shot generalization on the in-vivo RIS2017 and StereoMIS datasets.**

Received 16 April 2025; accepted 13 August 2025. Date of publication 1 September 2025; date of current version 10 September 2025. This article was recommended for publication by Associate Editor A. Kuntz and Editor J. Burgner-Kahrs upon evaluation of the reviewers' comments. This work was supported in part by EPSRC through the UCL Centre for Doctoral Training in Intelligent, Integrated Imaging in Healthcare (i4health) under Grant EP/S021930/1, in part by Human-centric Machine Intelligence to optimise Robotic Surgical Training under Grant EP/Z534754/1, in part by the Optical and Acoustic imaging for Surgical and Interventional Sciences under Grant UKRI145 projects, in part by UCL Research Excellence Scholarships Programme, in part by NIHR UCLH Biomedical Research Centre under Grant NIHR203328, and in part by the Department of Science, Innovation and Technology (DSIT) and the Royal Academy of Engineering through the Chair in Emerging Technologies programme. *(Corresponding authors: Xu Wang; Evangelos B. Mazomenos.)*

Xu Wang, Jialang Xu, and Evangelos B. Mazomenos are with UCL Hawkes Institute and the Department of Medical Physics and Biomedical Engineering, University College London, W1W 7TY London, U.K. (e-mail: xu.wang.23@ucl.ac.uk; jialang.xu.22@ucl.ac.uk; e.mazomenos@ucl.ac.uk).

Shuai Zhang and Danail Stoyanov are with UCL Hawkes Institute and the Department of Computer Science, University College London, W1W 7TY London, U.K. (e-mail: shuai.z@ucl.ac.uk; danail.stoyanov@ucl.ac.uk).

Baoru Huang is with the Department of Computer Science, University of Liverpool, L69 7ZX Liverpool, U.K. (e-mail: Baoru.Huang@liverpool.ac.uk).

Code is available at: https://github.com/MichaelWangGo/StereoMamba.git

This article has supplementary downloadable material available at https://doi.org/10.1109/LRA.2025.3604749, provided by the authors.

Digital Object Identifier 10.1109/LRA.2025.3604749

***Index Terms*—Stereo disparity estimation, robotic-assisted minimally invasive surgery, mamba.**

I. Introduction

STEREO endoscopes are routinely employed in robotic-assisted minimally invasive surgery (RAMIS) to visualize the internal anatomy, providing surgeons with depth perception for precise instrument manipulation [1]. Accurate, real-time, and robust disparity estimation from stereo video is a critical component in RAMIS, for understanding the geometry of the surgical scene and enabling downstream tasks such as preoperative image registration [2] and intraoperative navigation [3]. However, ensuring these capabilities especially in in-vivo environments, presents many challenges [4].

Stereo disparity estimation is fundamentally a matching task, where each pixel in the rectified left image searches along the epipolar line in the rectified right image to find the optimal corresponding pixel with the lowest matching cost. A typical stereo matching pipeline [7] can be split into three or four steps: feature extraction, cost volume construction, disparity estimation and/or refinement. For cost volume construction, many state-of-the-art (SOTA) approaches [5], [6], [8] first fix a maximum disparity range (typically 192 px is applied), and then compute the matching cost between the left and right feature maps for each disparity value starting from 0, creating a 4D cost volume (height × width × disparity × feature dimension). Then, the disparity value that minimizes the matching cost for each pixel pair is found within the 4D cost volume. However, relying solely on pixel-level features often leads to ambiguity, especially in RAMIS scenes that contain large textureless regions or repetitive patterns, where multiple pixel matches may appear equally optimal. To alleviate this issue, it is crucial to incorporate both fine-grained *local* features—captured from shallow Convolutional Neural Networks (CNNs)—and *global* features—representing broader spatial features across the image captured from deeper CNN layers—during the feature extraction stage [9].

PSMNet enhances stereo matching by leveraging global context information through spatial pyramid pooling (SPP) [9] and dilated convolutions, which extend pixel-level features to region-level representations across multiple scales. Another line of work focuses on improving cost volume construction effectiveness to further improve stereo matching accuracy. For example,

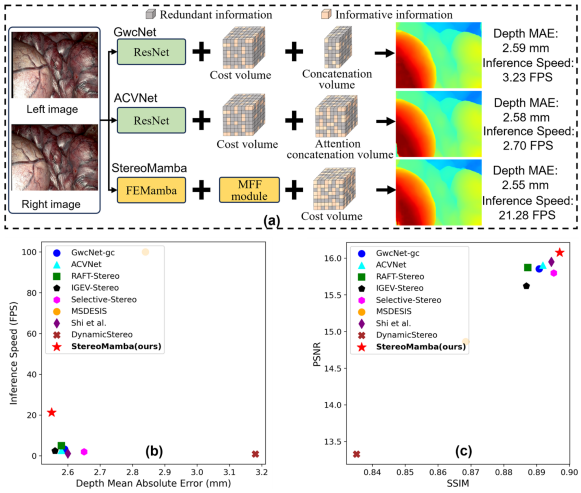

Fig. 1. Comparisons with SOTA disparity estimation methods. (a) shows the different methodologies between GwcNet [5], ACVNet [6], and our proposed StereoMamba. (b) illustrates the trade-off between accuracy and inference speed. (c) compares the similarity between synthesized right images (generated from left images and disparity maps on two unseen in-vivo datasets) and actual right images.

GwcNet [5] introduces a group-wise correlation mechanism to refine the cost volume, providing more precise matching features for disparity regression, shown in Fig. 1(a). However, these methods are limited by the receptive field of CNNs, which only extract features within single images. For stereo matching, establishing cross-image connections is essential for accurate correspondence retrieval [10]. ACVNet [6] also adopts CNNs for feature extraction, but introduces an attention concatenation volume that links left and right features during cost volume construction, which generates attention weights based on relevant cues in order to suppress redundant information and to enhance informative information in the cost volume, shown in Fig. 1(a). We propose that introducing these cross-image connections earlier, at the feature extraction stage, can improve stereo matching performance (shown in Fig. 1).

Transformer architectures could be a potential solution with their capacity to model long-range spatial dependencies [11]. The self-attention mechanism enables the extraction of global contextual features within a single image, while cross-attention facilitates correspondence between stereo image pairs. DynamicStereo incorporates self- and cross-attention to extract features across time and stereo pairs to maintain the temporal consistency of its predictions [12]. Cheng et al. [13] investigate Transformer integration within stereo matching pipelines [7], demonstrating that leveraging Transformers for feature representation learning and CNNs for cost aggregation leads to faster convergence, improved accuracy, and enhanced generalization. However, Transformer-based stereo matching methods often suffer from quadratic computational and memory complexity due to the Query-Key product [11]. This poses a significant limitation in real-time in-vivo applications, where both high accuracy and efficiency are critical. Striking a balance between estimation accuracy and model complexity remains a key challenge in advancing stereo disparity estimation.

Mamba [14], a selective State Space Model (SSM) [15] architecture originally proposed for natural language modelling, which combines the strengths of CNN's linear complexity and transformers' long-range spatial dependencies, is a promising alternative for sequence modelling. Inspired by these advancements, several SSM-based visual backbone networks, such as VMamba [16] and Vision Mamba [17], have been proposed for image classification [18] and segmentation [19]. However, no work has explored Mamba's unique capabilities for accurate, fast, and robust stereo disparity estimation in RAMIS, which remains an ongoing challenge in the field.

To this end, we propose StereoMamba, a novel end-to-end deep neural network designed for stereo disparity estimation in RAMIS. It features a powerful Feature Extraction Mamba (FE-Mamba) module that leverages both self-attention and cross-attention to enhance long-range spatial dependencies within and across stereo-pair images. To integrate multi-scale features from FE-Mamba to the stereo matching pipeline, we introduce a Multidimensional Feature Fusion (MFF) module, which seamlessly combines self-attentive and cross-attentive features. The fused features are then divided into multiple groups along the channel dimension, where each left feature group is cross-correlated with its corresponding right feature group across all disparity levels, generating group-wise correlation maps. These maps are subsequently aggregated to construct the final cost volume, which is processed by a disparity regression network to generate the final disparity maps.

Our main contributions are the following:

- *A novel feature extraction and fusion mechanism for stereo disparity estimation:* We propose FE-Mamba for feature extraction and MFF for multi-scale feature integration (Fig. 1(a)). Compared to existing SOTA methods GwcNet [5] and ACVNet [6], StereoMamba's feature extractor and fusion strategy provide more informative features for cost volume construction.
- *A balance between accuracy, robustness, and inference speed:* StereoMamba achieves a state-of-the-art EPE of 2.64 px and depth MAE of 2.55 mm on the SCARED benchmark while maintaining real-time inference at 21.28 FPS (Fig. 1(b)).
- *Strong zero-shot generalization on unseen surgical datasets:* By synthesizing right images using estimated disparity maps and original left images, StereoMamba achieves superior SSIM score of 0.8970, demonstrating high generalization ability (Fig. 1(c)).

II. Related Work

Iterative methods such as Recurrent All-Pairs Field Transforms (RAFT), build a 4D cost volume by computing correlations between all pixel pairs, and iteratively updating the flow field using a Gated Recurrent Unit (GRU)-based update operator [20]. RAFT-Stereo [21] extends this approach by employing multi-level Convolutional GRU (ConvGRU) to iteratively update the disparity field using local cost values retrieved from All-Pairs Correlations (APC). However, APCs lack global information and struggle with local ambiguities in challenging

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

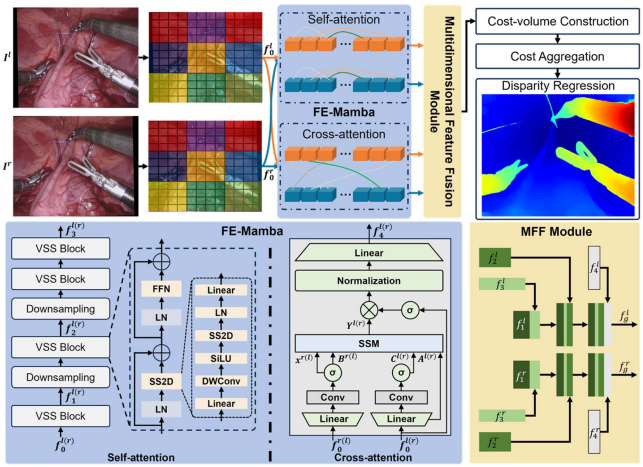

Fig. 2. The overall architecture of StereoMamba, which consists of the FE-Mamba module for feature extraction, the cost volume construction via the MFF module, and the cost volume aggregation for disparity estimation.

regions. In contrast, IGEV-Stereo introduces a module that encodes non-local geometry, contextual information, and local matching details, enhancing the effectiveness of each ConvGRU iteration [22]. This provides a better initial disparity map to the ConvGRUs, resulting in faster convergence. Selective-Stereo [8] outperforms both RAFT-Stereo and IGEV-Stereo, by proposing a Selective Recurrent Unit (SRU) and Contextual Spatial Attention (CSA) to adaptively fuse hidden disparity information at multiple frequencies for edge and smooth regions. However, these modules increase the sizes of convolutional kernels, leading to high memory and time costs. MSDESIS [1] proposes a multi-task network for surgical instrument segmentation and stereo disparity estimation, demonstrating that supervising the segmentation task enhances disparity estimation accuracy. Teacher-student methods have been also considered. Shi et al. [23] propose a dual-branch CNN-based teacher-student model, designed to further improve disparity estimation accuracy in surgical settings. This framework jointly trains the teacher-student network and a confidence network in a bidirectional semi-supervised manner, where each branch predicts disparity probability distributions, disparity values, and confidence maps. While this approach achieves high accuracy and robustness, it significantly increases computational demands.

III. Methodology

A. Network Architecture Overview

The StereoMamba architecture is illustrated in Fig. 2. It first applies a 2D convolution with a kernel size of 4 and a stride of 4 to downsample the rectified left and right images $I^{l(r)} \in \mathbb{R}^{H\times W\times 3}$, resulting in corresponding feature maps $f_0^{l(r)}\mathbb{R}^{\in \frac{H}{4}\times\frac{W}{4}\times C_0}$, where the superscript of $*^l$ and $*^r$ represent the left and right image or feature map. H, W and C_i denote the height, width and number of channels. To leverage global context information within and between the stereo image pair, we propose the FE-Mamba module to perform self- and cross-attention. Accordingly, the MFF module is proposed to effectively combine the self- and cross-attention features. We then follow [5] and split features into groups along the channel dimension to compute correlation maps for each group and construct a cost volume. Finally, a cost volume aggregation network is used to regress the disparity value.

B. Feature Extraction Mamba

The FE-Mamba module contains the components of self-attention and cross-attention. For self-attention, we leverage the four-way scanning strategy of VMamba [16] to process individual images. To model cross-attention between image pairs, we design a new visual component inspired by the Mamba2 architecture [24].

1) Self-Attention: VMamba consists of four stacks of Visual State-Space (VSS) blocks and downsampling layers that process multi-scale features. It utilizes 2D Selective Scan (SS2D) layers, depthwise convolutions (DWConv), SiLU activation, and feed-forward networks (FFN) to capture global features within image efficiently. These operations are interleaved with layer normalization (LN) and linear transformations, ensuring stable training and improved feature representations. Unlike the original VMamba, we eliminate the downsampling block from the final stack, since the low resolution feature output does not facilitate the subsequent upsampling process and leads to reduced stereo matching accuracy. The output features are denoted as $f_1^{l(r)} \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C_1}$, $f_2^{l(r)} \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times C_2}$, and $f_3^{l(r)} \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times C_3}$.

2) Cross-Attention: In Transformers, by far the most commonly used attention mechanism is softmax attention, defined as: $Y = softmax(QK^T)\cdot V$. A linear attention mechanism can be utilized to remove the softmax by folding it into a kernel feature map [25], and using the associativity of matrix multiplication to rewrite $(QK^T)\cdot V = Q\cdot(K^TV)$. In this case, a mask can be incorporated into the left-hand side as $(L\circ QK^T)\cdot V$, where L is the lower-triangular 1's matrix, and $\circ$ denotes element-wise product [25]. The final attention calculation can be written as:

$$Y = (L\circ QK^T)\cdot V \tag{1}$$

Considering the preliminaries of SSM that maps a 1-dimensional function or sequence $x(t)\in\mathbb{R}\mapsto y(t)\in\mathbb{R}$ through an implicit hidden state $h(t)\in\mathbb{R}^N$:

$$h_{t+1} = A_t h_t + B_t x_t; y_t = C_t h_t \tag{2}$$

by definition $h_0 = B_0 x_0$:

$$h_t = A_t\cdots A_1 B_0 x_0 + \cdots + A_t B_{t-1}x_{t-1} + B_t x_t \tag{3}$$

where we can denote $A^{\times}_{t:s} = A_t A_{t-1}\cdots A_{s+1}$ and $A^{\times}_{s:s} = I$. Then the output can be reformulated as:

$$y_t = \sum_{s=0}^{t} C_t^T A^{\times}_{t:s} B_s x_s \tag{4}$$

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Next we define a lower-triangular matrix $M_{t,s} = C_t^T A_{t:s}^{\times} B_s$, where t and s denote the row and column index of the matrix:

$$M = \begin{bmatrix} C_0^T B_0 & 0 & 0 & \cdots \\ C_1^T A_{1:0}^{\times} B_0 & C_1^T B_1 & 0 & \cdots \\ \vdots & \vdots & \ddots & \\ C_t^T A_{t:0}^{\times} B_0 & C_t^T A_{t:1}^{\times} B_1 & \cdots & C_t^T B_t^T \end{bmatrix} \tag{5}$$

then the output can be rewritten in the matrix format:

$$Y = Mx \tag{6}$$

Each element $M_{t,s}$ can be factorized as:

$$M_{t,s} = C_t^T A_{t:s}^{\times} B_s = (CB^T)_{t,s} \cdot A_{t:s}^{\times} \tag{7}$$

where $(CB^T)_{t,s} := C_t^T B_s$. This suggests that M can be expressed as the element-wise product between two matrices: a matrix CB^T whose (t,s)-th element is $C_t^T B_s$, and a lower-triangular matrix L whose (t,s)-th element is $A_{t:s}^{\times}$. Specifically, we define:

$$L_{t,s} = \begin{cases} A_{t:s}^{\times}, & t \geqslant s \\ 0, & t < s \end{cases} \tag{8}$$

Thus, we obtain a concise expression for the output:

$$Y = (L \circ CB^T) \cdot x \tag{9}$$

This equation is formally identical to (1), thus establishing a mathematically unified formulation of Mamba and Transformer architectures. For the sake of brevity, we write this form of (9) below as $Y = SSM(A, B, C, x)$.

Based on this, we design the cross-attention component of the FE-Mamba module, as shown in Fig. 2 (middle bottom). Given input feature maps $f_0^{l(r)}$, linear transformations are first applied to adjust their dimensions. Transformed features are then processed through convolutional layers, parameterized by $x^{r(l)}$, $B^{r(l)}$ and $C^{l(r)}$. The resulting features are then passed into the SSM, which captures long-range spatial dependencies between stereo images. The SSM generates intermediate outputs $Y^{l(r)}$:

$$Y^{l(r)} = SSM(A^{l(r)}, B^{r(l)}, C^{l(r)}, x^{r(l)}) \tag{10}$$

The final refined features $f_4^{l(r)}$ are obtained by applying a GeLU activation σ to $f_0^{l(r)}$, followed by a root mean squared normalization and a linear transformation.

$$f_4^{l(r)} = Linear\{RMSNorm\{Y^{l(r)}, \sigma(f_0^{l(r)})\}\} \tag{11}$$

C. Cost Volume With Multidimensional Feature Fusion

To effectively fuse self-attention features ($f_1^{l(r)}$, $f_2^{l(r)}$, $f_3^{l(r)}$) and cross-attention features ($f_4^{l(r)}$), we propose the MFF module. The feature map $f_3^{l(r)}$ from the final VSS block of the self-attention branch is up-sampled using a transposed convolutional layer with ReLU activation to match the dimensions of $f_1^{l(r)}$, and then concatenated with $f_1^{l(r)}$. The concatenated features are then passed through a transposed convolution and concatenated with $f_2^{l(r)}$, followed by the same convolution and activation. Finally, the merged self-attention features are concatenated with the cross-attention features $f_4^{l(r)}$ to obtain the multidimensional feature $f_g^{l(r)} \in \mathbb{R}^{\frac{H}{4}, \frac{W}{4}, C_g}$.

We then divide the multidimensional feature $f_g^{l(r)}$ along the channel dimension into groups and compute correlation maps for each group. All channels are evenly divided into N_g groups, therefore each feature group has C_g/N_g channels. The group-wise correlation is computed as:

$$C_{gwc}(d, x, y, i) = \frac{1}{C_g/N_g} \left\langle f_g^{l,i}(x, y), f_g^{r,i}(x, y) \right\rangle \tag{12}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product, and $f_g^{l,i}$, $f_g^{r,i}$ represent the i-th feature group of the left and right multidimensional features, respectively.

D. Loss Function

Four outputs are obtained in the cost aggregation. For each output, two 3D convolutions are employed to produce a 1-channel volume, which we then up-sample and convert into a probability volume using softmax along the disparity dimension. For each pixel, we have a D_{max}-length vector which contains the probability p for all disparity levels. Finally, the predicted disparity value is computed by the soft-argmin function,

$$\hat{d} = \sum_{k=0}^{D_{max}-1} k \cdot p_k \tag{13}$$

where k and p_k denote the candidate disparity and corresponding probability. The four predicted disparity maps ($\hat{d_0}$, $\hat{d_1}$, $\hat{d_2}$, and $\hat{d_3}$) are used to formulate the overall loss as:

$$Loss = \sum_{i=0}^{3} w_i \cdot L_{smooth_{L1}}(\hat{d_i}, d) \tag{14}$$

where w_i denotes the coefficients for the ith disparity prediction and d denotes the ground-truth disparity map. The smooth L1 loss $L_{smooth_{L1}}$ is defined as:

$$L_{smooth_{L1}}(x, y) = \begin{cases} 0.5(x-y)^2, & \text{if } |x-y| < 1 \\ |x-y| - 0.5, & \text{if } otherwise \end{cases} \tag{15}$$

IV. Model Development

A. Datasets and Evaluation Metrics

SceneFlow [26] is a synthetic stereo collection that includes three subsets: Flyingthings3D, Driving, and Monkaa. The resolution is 960 × 540 and dense disparity maps are provided as ground truth. Following the original splitting strategy, we use the cleanpass category, selecting 35,454 image pairs for training and 4,370 for testing. **SCARED** is a surgical video dataset, from the MICCAI 2019 Endovis challenge, featuring depth maps of porcine abdominal anatomy captured using a da Vinci Xi endoscope and structured light projectors [27]. It includes 7 training subsets and 2 testing subsets, each containing 4 videos and one image pair at 1280 × 1024 resolution. Due to calibration errors in subsets 4 and 5 and synchronization issues between RGB videos and depth maps, we use only the first keyframes from subsets 1, 2, 3, 6, and 7, yielding 25 training

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

image pairs. Evaluation is conducted on 2 test subsets, each with 5 keyframes (4 video sequences and 1 image pair), totalling 5909 image pairs. **Robotics Instrument Segmentation (RIS_2017) Dataset** [28] is released as part of the MICCAI 2017 Robotic Instrument Segmentation Challenge. This dataset is generated from 10 abdominal porcine operations recorded using the da Vinci Xi system and provides rectified stereo frame data with a resolution of 1280 × 1024. The dataset is sampled at a rate of 1 Hz, yielding 3,000 stereo image pairs. **StereoMIS** [29] is a dataset used for Simultaneous Localization and Mapping (SLAM) in endoscopic surgery, without any disparity or depth ground truth. This dataset is collected using the da Vinci Xi surgical robot. It consists of 10 stereo video sequences recorded at 1280 × 1024 resolution. We extract one frame per ten frames, yielding 12,180 stereo image pairs. RIS_2017 and StereoMIS do no have disparity ground truth and are used to evaluate zero-shot generalization.

Evaluation Metrics: Following [1], [5], we use established disparity evaluation metrics: End-Point Error (EPE), Bad2, Bad3, Bad5, depth MAE, and inference speed. EPE represents the mean absolute error between the ground truth and predicted disparity values. Bad2, Bad3 and Bad5 indicates the percentage of pixels where the estimated disparity deviates by more than 2 pixels, 3 pixels, 5 pixels from the ground truth. Depth MAE, measured in millimeters, quantifies the mean absolute error in depth estimation. Inference speed is measured in frames per second (FPS). In all performance metrics, lower values indicate better results. To evaluate the generalization ability on the two datasets without ground truth, we adopt [30] and warp left images using disparity maps to synthesize right images, then compare them with actual right images. The similarity is assessed using three widely used metrics: Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), and Learned Perceptual Image Patch Similarity (LPIPS).

B. Implementation Details

All experiments are implemented in PyTorch on a single Nvidia RTX A6000 GPU with 48 GB of memory. Initially, StereoMamba is pre-trained on the SceneFlow (cleanpass) dataset for 40 epochs with a batch size of 14. We adopt the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 1e-4. We employ a one-cycle learning rate policy with a maximum learning rate of 2e-4, keeping the other hyperparameters at PyTorch's default values. Data augmentation includes random crops of size 512 × 256 and color normalization based on each dataset's statistics. The maximum disparity value is $D_{max} = 192$. The coefficients of four outputs are set as $w_0 = 0.5$, $w_1 = 0.5$, $w_2 = 0.7$, $w_3 = 1.0$ following [5]. Subsequently, the model is fine-tuned on the SCARED dataset with a constant learning rate of 1e-3 for 150 epochs.

V. Experiments and Results

A. Comparison With SOTA on the SCARED Benchmark

Following the recommended evaluation protocol from [27], the evaluation is performed on all frames except those in which more than 90% of the ground truth disparity maps are empty. Seven SOTA methods including, the baseline method Gwc-Net [5], one cost volume-optimization method ACVNet [6], three iterative-optimization methods RAFT-Stereo [21], IGEV-Stereo [22] and Selective-Stereo [8], one multi-tasking method MSDESIS [1], one semi-supervised teacher-student method Shi et al. [23] and one Transformer-based method DynamicStereo [12] are chosen for comparison. For methods that follow the standard pre-training on SceneFlow and fine-tuning pipeline [1], [5], [6], [8], [12], [21], [22], we fine-tune them on the same SCARED dataset as ours to ensure a fair comparison. Since Shi et al. [23] only provides inference code and weights trained on SCARED, we use the released model to directly generate disparity maps. DynamicStere [12] which is specifically designed for sequential inputs, cannot be applied in K4, since each subset contains only single-frame stereo pairs.

Table I lists comparative results of StereoMamba against the SOTA approaches. Our StereoMamba outperforms all competing methods on K2 and K3 videos of Test Subset 1, achieving depth MAEs of 1.46 mm and 2.03 mm, respectively. For Test Subset 2, StereoMamba ranks second-best on K0, K1, K2, and K3 videos, with performance differences ranging from just 0.01 mm to 0.05 mm compared to the top methods [5], [6], [21]. The results demonstrate that StereoMamba achieves comparable or even superior disparity estimation accuracy compared to other SOTA methods.

In summary, StereoMamba outperforms all competing methods on the entire SCARED dataset in terms of both mean depth MAE (2.55 mm) and EPE (2.64 px). It also ranks as second-best on Bad2 (41.49%) and Bad3 (26.99%), trailing the top-performing ACVNet [6] by only 0.83% and 0.44%, respectively. For Bad5, StereoMamba reports a value of 13.88%, just 0.06% behind the best-performing method IGEV-Stereo [22]. Although other methods perform well on individual keyframes, their overall performance on the entire dataset is inferior to StereoMamba. This consistently strong performance across multiple evaluation metrics highlights StereoMamba's robustness in handling the challenging areas in SCARED, including specular reflections and textureless regions that are particularly difficult for stereo matching.

Fig. 3 shows qualitative results in challenging areas. Notably, compared to the other methods, StereoMamba demonstrates superior performance at image edges and in dark regions, as highlighted by the red dashed-lined boxes. StereoMamba maintains low depth MAE, Bad2, Bad3, Bad5 and EPE, indicating its reliability, which is particularly important for tasks such as SLAM and 3D reconstruction where accurate boundary disparity ensures stable tracking.

More importantly, StereoMamba delivers real-time performance with an inference speed at 21.28 FPS, significantly outperforming cost volume methods [5], [6], iterative-optimization methods [8], [21], [22], and the semi-supervised teacher-student method [23], which operate only at 1.08 to 5.00 FPS. In particular, it achieves a substantial speedup over the Transformer-based DynamicStereo [12], whose inference speed is limited to 0.86 FPS. Although MSDESIS [1] achieves real-time inference, it suffers from poor disparity estimation accuracy. Overall, StereoMamba exhibits strong robustness while effectively

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE I
EVALUATION ON SCARED TEST SUBSETS

Methods	Test subset 1 (Depth MAE: mm ↓)					Test subset 2 (Depth MAE: mm ↓)					mean Depth MAE (mm)↓	Bad2 (%)↓	Bad3 (%)↓	Bad5 (%)↓	EPE (px)↓	Inference speed (FPS)↑
	K0	K1	K2	K3	K4	K0	K1	K2	K3	K4						
GwcNet-gc [5]	8.60	2.63	1.49	2.06	0.66	4.25	1.00	3.41	**1.44**	0.35	2.59	41.55	27.28	14.07	2.67	3.23
ACVNet [6]	8.56	2.59	1.49	2.08	0.68	4.23	0.96	3.46	**1.44**	**0.33**	2.58	**40.66**	**26.50**	13.85	2.70	2.70
RAFT-Stereo [21]	8.45	2.51	1.75	**2.03**	0.73	**4.14**	**0.89**	**3.06**	1.52	0.68	2.58	44.91	29.80	14.19	2.83	5.00
IGEV-Stereo [22]	8.27	2.35	1.73	2.18	**0.56**	4.41	0.97	3.22	1.54	0.37	2.56	42.15	27.28	**13.82**	2.65	2.44
Selective-Stereo [8]	8.44	2.44	1.62	2.27	0.70	4.30	0.94	3.34	1.78	0.66	2.65	43.48	28.36	14.17	2.81	1.96
MSDESIS [1]	8.42	2.59	2.05	3.02	1.01	4.68	1.18	3.34	1.61	0.46	2.84	44.71	29.81	16.14	3.06	**100.00**
Shi *et al.* [23]	**7.61**	**2.10**	1.97	2.66	0.65	4.75	1.18	2.96	1.71	0.36	2.60	44.41	28.54	14.28	2.74	1.08
DynamicStereo [12]	8.64	2.50	1.66	1.95	-	4.23	1.03	3.76	1.66	-	3.18	47.95	32.38	16.96	2.99	0.86
StereoMamba (Ours)	8.57	2.61	**1.46**	**2.03**	0.77	4.19	0.92	3.07	1.46	0.41	**2.55**	41.49	26.99	13.88	**2.64**	21.28

Kn represents the n*th* keyframe of each subset.The table reports depth MAE for each keyframe, along with the mean depth MAE, Bad2, Bad3, Bad5, and EPE across all test keyframes. Bold: Best, Underline: Second-best.

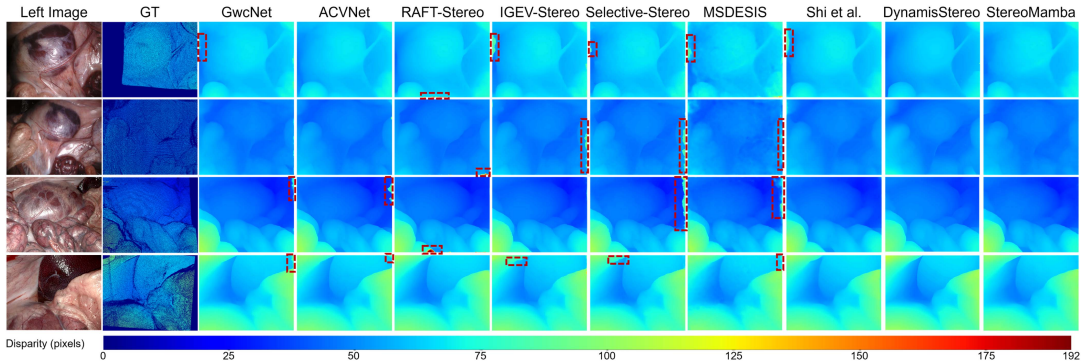

Fig. 3. Qualitative results on SCARED. The first column indicates the rectified left image, the second column indicates the ground truth disparity, and the other columns are estimated disparity maps from GwcNet [5], ACVNet [6], RAFT-Stereo [21], IGEV-Stereo [22], Selective-Stereo [8], MSDESIS [1], Shi et al. [23], DynamicStereo [12] and our StereoMamba. All methods share the same disparity colorbar, ranging from 0 to 192 pixels. Additional visualization results are provided in the supplementary material.

TABLE II
THE EVALUATION OF GENERALIZATION ABILITY ON STEREOMIS AND RIS_2017 DATASETS

	StereoMIS			RIS_2017			Average		
	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓
GwcNet-gc [5]	0.9120	17.1222	0.2834	0.8697	14.5847	0.3539	0.8908	15.8534	0.3187
ACVNet [6]	0.9141	**17.3140**	**0.2770**	0.8696	14.4944	0.3428	0.8919	15.9042	0.3099
RAFT-Stereo [21]	0.9108	17.1873	0.2781	0.8637	14.5558	**0.3303**	0.8873	15.8715	**0.3042**
IGEV-Stereo [22]	0.9096	16.9007	0.2823	0.8641	14.3403	0.3440	0.8869	15.6205	0.3131
Selective-Stereo [8]	0.9110	16.8081	0.2888	**0.8793**	14.7840	0.3592	0.8952	15.7961	0.3240
MSDESIS [1]	0.8842	15.6194	0.3263	0.8526	14.1028	0.3699	0.8684	14.8611	0.3481
Shi *et al.* [23]	0.9134	17.2028	0.2818	0.8755	14.6991	0.3353	0.8945	15.9509	0.3086
DynamicStereo [12]	0.8417	13.2468	0.3948	0.8287	13.4070	0.3453	0.8352	13.3269	0.3701
StereoMamba (Ours)	**0.9149**	17.3054	0.2786	0.8790	**14.8468**	0.3431	**0.8970**	**16.0761**	0.3109

balancing disparity estimation accuracy and inference speed, achieving the optimum trade-off for real-world RAMIS applications.

B. Zero-Shot Generalization Results

Considering the high variability in surgical scenes due to different patient anatomy, hardware utilized and the various surgical applications of RAMIS (e.g. urology, gynaecology), the generalization ability of stereo disparity estimation models is a key performance indicator. We thus evaluate the zero-shot generalization performance of StereoMamba and other methods in unseen real-world, in-vivo surgical scenes. Following a zero-shot setting, all methods are trained on SceneFlow, fine-tuned on SCARED, and directly tested on the RIS_2017 and StereoMIS datasets.

As shown in Table II, our approach achieves comparable or superior performance against SOTA methods across both in-vivo datasets. On StereoMIS, StereoMamba attains the best SSIM score (0.9149), the second-best PSNR score (17.3054), and an LPIPS score of 0.2786, just 0.0016 higher than the best result. In RIS_2017, StereoMamba achieves the best PSNR score (14.8468), the second-best SSIM score (0.8790), and an LPIPS score of 0.3431, only 0.0128 higher than the top method. On average, StereoMamba achieves the best SSIM (0.8970) and PSNR (16.0761), with LPIPS being [21] only 0.0067 higher than the best method. Evidently, StereoMamba demonstrates strong generalization capability in generating reliable disparity maps on unseen in-vivo datasets.

Example results are presented in Fig. 4, with notable disparity estimation errors highlighted in dashed red boxes. GwcNet [5] struggles with dark regions and specular reflections from instruments. ACVNet [6] improves performance in dark areas but still suffers from specular reflections. RAFT-Stereo [21] handles these challenges better but loses some instrument details on the RIS_2017 dataset. IGEV-Stereo [22] and Selective-Stereo [8]

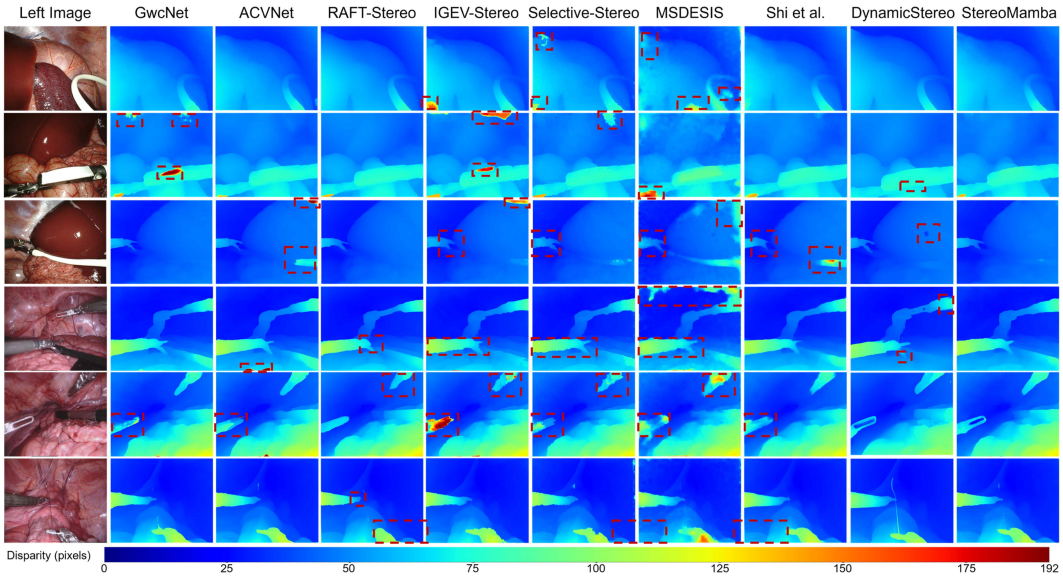

Fig. 4. Qualitative results on StereoMIS (first 3 rows) and RIS_2017 (last 3 rows). The first column indicates the rectified left image, the other columns are estimated disparity maps from GwcNet [5], ACVNet [6], RAFT-Stereo [21], IGEV-Stereo [22], Selective-Stereo [8], MSDESIS [1], Shi et al. [23], DynamicStereo [12] and our StereoMamba. All methods share the same disparity colorbar, ranging from 0 to 192 pixels. Additional visualization results are provided in supplementary material.

TABLE III
ABLATION STUDY RESULTS OF PROPOSED NETWORKS ON THE CLEANPASS OF SCENEFLOW DATASET

Method	ResNet	Transformer	FE-Mamba	MFF	EPE (px) ↓	Bad2 (%) ↓	Bad3 (%)↓	Bad5 (%)↓	Inference speed (FPS)↑
GwcNet [5]	√				0.7691	4.37	3.30	2.25	4.76
GwcNet*		√			1.2417	6.52	4.85	3.48	0.56
StereoMamba-base			√		**0.6044**	3.57	2.26	1.80	**27.78**
StereoMamba			√	√	**0.6044**	**3.48**	**2.25**	**1.76**	27.03

All metrics are for 960×540 inputs on a single Nvidia RTX A6000 GPU.

struggle with image edges, dark regions, and specular reflections, leading to unsmooth instrument boundaries. MSDESIS [1] performs the worst, producing disparity maps with significant noise, while Shi et al. [23] shows slight improvement but remains affected by specular reflections. In contrast, StereoMamba demonstrates strong robustness against these challenges, highlighting its superiority for zero-shot generalization compared to other methods.

C. Ablation Study

To verify the effectiveness of our proposed modules, we take GwcNet [5] as the baseline and replace its ResNet backbone with our FE-Mamba and MFF modules. As shown in Table III, replacing ResNet with FE-Mamba alone (StereoMamba-base) leads to notable improvements: EPE is reduced by 0.1651 px, Bad2 by 0.8%, Bad3 by 1.04%, and Bad5 by 0.45%, while inference speed significantly increases from 4.76 FPS to 27.78 FPS, demonstrating the efficiency and effectiveness of FE-Mamba. With the addition of the MFF module, the full StereoMamba model achieves further improvements, reducing Bad2 by an additional 0.09%, Bad3 by 0.01%, and Bad5 by 0.04%. Although the inference speed slightly decreases to 27.03 FPS, the model runs at 21.28 FPS on the 1280 × 1024 image pairs of SCARED, the trade-off between inference speed and accuracy is acceptable. The inference speed of 21.28 FPS is sufficient for real-time depth estimation during surgery, and the marginal improvement in speed beyond this threshold offers limited practical benefit. From this perspective, we believe that the modest reduction in FPS is fully justified by the improved accuracy. In addition, we implemented a Transformer-based feature extractor following the design concept of GwcNet (denoted as GwcNet* in Table III). Specifically, the Transformer feature extractor first extracts features at three different scales and then concatenates them. However, its inference speed is only 0.56 FPS, which further underscores the efficiency advantages of the proposed FE-Mamba and MFF modules.

VI. DISCUSSION

A. Design Rationale of FE-Mamba and MFF

The FE-Mamba module integrates self- and cross-attention at the feature extraction stage to address two complementary demands in stereo matching: (1) self-attention enhances intra-image semantic context and reduces local ambiguities, and (2) cross-attention explicitly enforces geometric consistency across views. The MFF module adopts a hierarchical fusion strategy rather than simple concatenation. By progressively upsampling and merging multi-scale self-attention features before integrating them with cross-attention, MFF preserves fine-grained spatial details, incorporates high-level context, and aligns features to the resolution required for effective cost volume computation.

B. Statistical Validation of Performance Gains

Although the mean differences between StereoMamba and competing methods appear small, frame-by-frame paired t-tests confirm that these improvements are statistically significant. Specifically: EPE vs. IGEV-Stereo: $t = -10.07$, $p = 1.34 \times 10^{-23}$; Depth MAE vs. IGEV-Stereo: $t = -5.78$, $p = 7.98 \times 10^{-9}$; SSIM vs. Selective-Stereo: $t = 9.67$, $p = 1.25 \times 10^{-21}$; PSNR vs. Shi et al.: $t = 10.27$, $p = 4.14 \times 10^{-24}$. The

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

extremely low p-values indicate that the observed gains are unlikely to be due to chance, confirming that StereoMamba consistently outperforms prior models.

C. Failure Cases

Although StereoMamba achieves high overall accuracy, occasional matching failures can occur in challenging scenarios. In particular, frames with very low texture, smoke, or extreme scene distances from the camera may lead to local mismatches or artifacts in the disparity maps, which are common limitations in stereo matching. Analysing these failure modes provides insights into situations where the method should be applied with caution and highlights directions for future improvement.

VII. Conclusion

In this letter, we propose StereoMamba, the first method to explore SSM for disparity estimation in RAMIS. We design a specialized FE-Mamba module to perform both self-attention and cross-attention within and cross stereo images, effectively encoding long-range spatial features, which are then seamlessly integrated using our novel MFF module. The fused multidimensional features are processed by a group-wise correlation-based decoder to generate the final disparity map. On the SCARED benchmark, StereoMamba achieves SOTA performance with an EPE of 2.64 px and a Depth MAE of 2.55 mm. It also delivers competitive results on Bad2 (41.49%), Bad3 (26.99%) and Bad5 (13.88%), while maintaining a real-time inference speed of 21.28 FPS for 1280×1024 image pairs. Compared to existing methods, it produces smooth and stable disparity estimations, even in challenging regions such as specular reflections and textureless areas. This balance between accuracy, robustness and inference speed makes StereoMamba well-suited for real-world deployment. Additionally, StereoMamba demonstrates strong zero-shot generalization on two unseen in-vivo datasets (RIS_2017, StereoMIS), achieving an SSIM of 0.8970, PSNR of 16.0761, and LPIPS of 0.3109 when comparing synthesized right images with the actual ones.

References

[1] D. Psychogyios, E. Mazomenos, F. Vasconcelos, and D. Stoyanov, "MS-DESIS: Multitask stereo disparity estimation and surgical instrument segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 11, pp. 3218–3230, Nov. 2022.

[2] Z. Chen et al., "FRSR: Framework for real-time scene reconstruction in robot-assisted minimally invasive surgery," *Comput. Biol. Med.*, vol. 163, 2023, Art. no. 107121.

[3] X. Feng, X. Zhang, X. Shi, L. Li, and S. Wang, "ST-ITEF: Spatio-temporal intraoperative task estimating framework to recognize surgical phase and predict instrument path based on multi-object tracking in keratoplasty," *Med. Image Anal.*, vol. 91, 2024, Art. no. 103026.

[4] Z. Chen et al., "Spatio-temporal layers based intra-operative stereo depth estimation network via hierarchical prediction and progressive training," *Comput. Methods Programs Biomed.*, vol. 244, 2024, Art. no. 107937.

[5] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3273–3282.

[6] G. Xu, J. Cheng, P. Guo, and X. Yang, "Attention concatenation volume for accurate and efficient stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12981–12990.

[7] X. Cheng et al., "Hierarchical neural architecture search for deep stereo matching," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 22158–22169.

[8] X. Wang, G. Xu, H. Jia, and X. Yang, "Selective-Stereo: Adaptive frequency information selection for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 19701–19710.

[9] J. -R. Chang and Y. -S. Chen, "Pyramid stereo matching network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5410–5418.

[10] Y. Ding et al., "TransMVSNet: Global context-aware multi-view stereo network with transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8585–8594.

[11] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–21.

[12] N. Karaev, I. Rocco, B. Graham, N. Neverova, A. Vedaldi, and C. Rupprecht, "DynamicStereo: Consistent dynamic depth from stereo videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 13229–13239.

[13] X. Cheng, Y. Zhong, M. Harandi, T. Drummond, Z. Wang, and Z. Ge, "Deep laparoscopic stereo matching with transformers," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2022, pp. 464–474.

[14] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *Proc. Int. Conf. Learn. Representations*, 2024, pp. 1–27.

[15] A. Gu, K. Goel, and C. Re, "Efficiently modeling long sequences with structured state spaces," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–27.

[16] Y. Liu et al., "VMamba: Visual state space model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, vol. 37, pp. 103031–103063.

[17] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," in *Proc. Int. Conf. Mach. Learn.*, 2024, vol. 235, pp. 62429–62442.

[18] Y. Li, Y. Luo, L. Zhang, Z. Wang, and B. Du, "MambaHSI: Spatial–spectral mamba for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5524216.

[19] C. Ma and Z. Wang, "Semi-Mamba-UNet: Pixel-level contrastive and cross-supervised visual Mamba-based UNet for semi-supervised medical image segmentation," *Knowl. Based Syst.*, vol. 300, 2024, Art. no. 112203.

[20] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 402–419.

[21] L. Lipson, Z. Teed, and J. Deng, "RAFT-stereo: Multilevel recurrent field transforms for stereo matching," in *Proc. Int. Conf. 3D Vis.*, 2021, pp. 218–227.

[22] G. Xu, X. Wang, X. Ding, and X. Yang, "Iterative geometry encoding volume for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 21919–21928.

[23] 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 Trans. Med. Imag.*, vol. 42, no. 11, pp. 3269–3282, Nov. 2023.

[24] T. Dao and A. Gu, "Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality," in *Proc. Int. Conf. Mach. Learn.*, 2024, pp. 31788–31812.

[25] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are RNNs: Fast autoregressive transformers with linear attention," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 5156–5165.

[26] N. Mayer et al., "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4040–4048.

[27] M. Allan et al., "Stereo correspondence and reconstruction of endoscopic data challenge," 2021, *arXiv:2101.01133*.

[28] M. Allan et al., "2017 robotic instrument segmentation challenge," 2019, *arXiv:1902.06426*.

[29] M. Hayoz et al., "Learning how to robustly estimate camera pose in endoscopic videos," *Int. J. Comput. Assist. Radiol.*, vol. 18, no. 7, pp. 1185–1192, 2023.

[30] B. Yan, C. Ma, B. Bare, W. Tan, and S. C. Hoi, "Disparity-aware domain adaptation in stereo image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13179–13187.