Misaligned 3D Texture Optimization in MIS Utilizing Generative Framework

Jieyu Zheng[1], Xiaojian Li[1] *, Hangjie Mo[1], Ling Li[1], and Xiang Ma[1]

Key Laboratory of Process Optimization and Intelligent Decision-making, National-Local Joint Engineering Research Center for Intelligent Decision-making and Information System Technology, School of Management, Hefei University of Technology, Anhui, China
zjylearn@mail.hfut.edu.cn, lixj90@hfut.edu.cn

Abstract. Three-dimensional reconstruction of the surgical area based on intraoperative laparoscopic videos can restore 2D information to 3D space, providing a solid technical foundation for many applications in computer-assisted surgery. SLAM methods often suffer from imperfect pose estimation and tissue motion, leading to the loss of original texture information. On the other hand, methods like Neural Radiance Fields and 3D Gaussian Split require offline processing and lack generalization capabilities. To overcome these limitations, we explore a texture optimization method that generates high resolution and continuous texture. It designs a mechanism for transforming 3D point clouds into 2D texture space and utilizes a generative network architecture to design 2D registration and image fusion modules. Experimental results and comparisons with state-of-the-art techniques demonstrate the effectiveness of this method in preserving the high-fidelity texture.

Keywords: 3D Texture optimization · Misaligned information fusion · Generative framework · High-fidelity texture· Registration.

1 Introduction

The 3D reconstruction of the surgical area finds wide applications in medical domains such as surgical navigation [7, 19], Augmented Reality (AR) systems [16, 4], and remote surgical guidance [2]. Simultaneous Localization and Mapping (SLAM) methods combine deep learning to achieve incremental dense reconstruction, demonstrating potential for real-time application in laparoscopic surgery [12, 20, 9, 7, 13]. However, influenced by imperfect pose estimation results and subtle tissue movements, simple point cloud overlay inevitably leads to the loss of original texture information in the reconstructed internal 3D structure. Since the input intraoperative laparoscopic video can be seen as continuous dense observations of the surgical area, each frame carries more valid information. Therefore, the current limitation of SLAM algorithms in faithfully reproducing the consistent texture is an unreasonable phenomenon.

* Corresponding author.

Recently, Neural Radiance Field (NeRF) [15, 14] and 3D Gaussian Splitting (3DGS) [5, 8] have demonstrated their potential in generating high-quality images and reconstructing geometric structures. They are trained on extensive photo collections of the entire scene utilizing viewpoint constraints. However, the limitation of both NeRF and 3DGS is that they are offline processes and lack generalization. They require the pre-calibration of pose information for each image or the initialization of sparse point clouds using tools like COLMAP[11]. In real surgical environments, this pre-calibration step is impractical because the scene is dynamic and constantly changing. Moreover, all the steps need to be repeated for a new surgical scene.

The migration of SLAM technology to endoscopes or laparoscopes has become a popular topic. Previous methods have explored the accuracy of depth estimation and pose estimation from various perspectives, but typically there are only three approaches commonly used for point cloud fusion. MIS-SLAM, EndoMotion, and SAGE-SLAM[12, 9, 7] adopt volume and surfel-based fusion methods to weight and sum the textures of point clouds. EMDQ-SLAM[20] uses the multi-band blending (MBB) to generate a smooth transition effect. BDIS-SLAM[13] simplifies dense frame-wise point cloud fusion by replacing overlapping regions with the latest frame. All methods suffer from a decrease in the resolution of 3D textures when handling inaccurate poses and outlier noise, resulting in a visually blurry effect or loss of original texture details. A similar field is 2D SLAM[21, 6] based on laparoscopic images, but they only perform image mosaicking in the 2D image coordinate system, lacking the mapping relationship with 3D space. As a result, they cannot utilize the reverse optimization of point cloud texture using the stitched 2D texture.

In this paper, we propose a novel texture optimization method, which could merge RGB point clouds from two frames into a unified point cloud with high-definition and continuous textures. By inputting RGB images for each frame, existing algorithms can be used to obtain depth and pose with estimation errors. The key challenge lies in effectively overlaying information using 3D misalignment data. We transform the 3D point clouds of two frames into a unified coordinate system, then use camera projection and regression algorithms to obtain 2D imaging. Finally, we construct a generative model with registration and fusion module to fuse misaligned information from the two frames, so as to directly generate the seamless and high-fidelity texture for point cloud.

2 Methodology

2.1 2D Mapping

Due to the disorder and discreteness of 3D point cloud data, directly optimizing the texture of 3D point clouds often involves complex methods and extensive computations. In the SLAM framework, assuming we have RGB images, depth maps, and camera poses between consecutive frames, the point cloud information for each frame is derived from the 2D domain. Therefore, we aim to drive 3D texture optimization through the fusion of multi-frame 2D information. Initially,

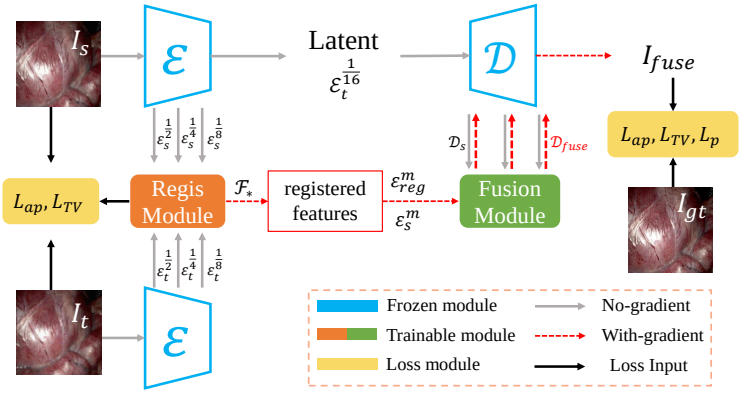

Fig. 1. The designed architecture based on KL-Reg VQ-GAN.

using estimated camera poses, we transform the point cloud P_i of frame i to frame j as $P_{i,j}$, then obtain discrete pixels p_{pixel} in the 2D pixel coordinate system using the camera model. These fractional values cannot directly form an image. So, based on p_{pixel}, Delaunay triangulation is performed. Then, for each integer pixel position within different triangles, interpolation calculates the corresponding texture information, thus obtaining the observed image $I_{i,j}$ of point cloud $P_{i,j}$. This functionality can be directly implemented in SciPy using the **griddata** function. Finally, $I_{i,j}$ and the captured RGB image I_j are used as input to the texture optimization network.

2.2 Registration Module

The VQ-GAN[3] employs a two-stage approach to image generation, which involves compression and regeneration. The encoder can compress images into compact latent spaces while providing multi-scale features with more effective information for better registration. The decoder can generate high-resolution images from limited information, providing a coarse-to-fine pattern for texture fusion and refine. Therefore, we construct the texture optimization network by combining the designed *Registration Module* and *Fusion Module* based on the pretrained KL-Reg VQ-GAN[10], as shown in Fig.1.

Assuming the input image is $I \in \mathbb{R}^{H\times W\times 3}$, the KL-Reg VQ-GAN encoder compresses it into a latent code $\mathcal{E}^{\frac{1}{16}}$, while providing multi-scale deep features $(\mathcal{E}^1, \mathcal{E}^{\frac{1}{2}}, \mathcal{E}^{\frac{1}{4}}, \mathcal{E}^{\frac{1}{8}})$. Here, the numbers denote the downsampling ratios. Since the rotation and translation between the two texture frames are unknown, a coarse-to-fine technique is employed to accurately predict the texture correspondence.

Coarse-scale Transformer Given source frame I_s and target frame I_t, we build a transformer architecture shown in Fig.2(a), to compute global correlation at $\mathcal{E}_s^{\frac{1}{8}}, \mathcal{E}_t^{\frac{1}{8}} \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times C_e}$, where C_e denotes the number of channels. To account for the texture similarity of laparoscopic images and the permutation invariance

of attention mechanisms, calculating 2D position encodings assist the network in recognizing the position and order of each feature, thereby eliminating potential matching ambiguities. Then the input 3D feature maps $\mathcal{E}_s^{\frac{1}{8}}, \mathcal{E}_t^{\frac{1}{8}}$ are element-wise added with 2D positional encoding and flattened into 2D feature maps $e_I \in \mathbb{R}^{\frac{HW}{64} \times C_e}$ before being fed into the transformer.

The transformer architecture is composed of multiple stacked self-attention and cross-attention modules. For each attention head h, fully connected layers are used to compute the query matrix $\mathcal{Q}_h$, key matrix $\mathcal{K}_h$, and value matrix $\mathcal{V}_h$ from the input features. In the self-attention module, $\mathcal{Q}_h, \mathcal{K}_h, \mathcal{V}_h$ are all derived from the same input, while in the cross-attention module, $\mathcal{Q}_h$ is derived from the source image (target image), and $\mathcal{K}_h, \mathcal{V}_h$ are derived from the target image (source image), which are shown in Fig.2(a). The attention map $\mathcal{A}_h \in \mathbb{R}^{\frac{HW}{64} \times \frac{HW}{64}}$ in each head is computed via the scaled matrix multiplication and softmax function. The output feature map $\mathcal{V}_h$ is then computed by:

$$\mathcal{V}_{\mathcal{O}} = W_{\mathcal{O}} \text{Concat}(\mathcal{A}_1 \mathcal{V}_1, \mathcal{A}_2 \mathcal{V}_2, \cdots, \mathcal{A}_H \mathcal{V}_H) + b_{\mathcal{O}}, \tag{1}$$

where H means the number of attention heads, $W_{\mathcal{O}} \in \mathbb{R}^{C_e \times C_e}$ and $b_{\mathcal{O}} \in \mathbb{R}^{C_e}$ are learnable parameters of a fully connected layer. Then e_I is added directly to $\mathcal{V}_{\mathcal{O}}$ through residual connection, and the final features $e_{\mathcal{O}} \in \mathbb{R}^{\frac{HW}{64} \times C_e}$ for the next layer are obtained by applying Layer Normalization and fully connected layer. The specific network structure is depicted in Fig.2(b).

Another pathway in cross-attention module involves deriving the optical flow between two images from the multi-head attention. The global attention map $\mathcal{A} \in \mathbb{R}^{\frac{HW}{64} \times \frac{HW}{64}}$ is defined as:

$$\mathcal{A} = \text{softmax}\left(\frac{\mathcal{Q}_1 \mathcal{K}_1^T}{\sqrt{C_h}} + \frac{\mathcal{Q}_2 \mathcal{K}_2^T}{\sqrt{C_h}} + \cdots + \frac{\mathcal{Q}_H \mathcal{K}_H^T}{\sqrt{C_h}}\right). \tag{2}$$

The optical flow $\mathcal{F}^{\frac{1}{8}} \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 2}$ is then computed by the weighted sum of the attention map $\mathcal{A}$ and the corresponding position:

$$\mathcal{F}_{i,j}^{\frac{1}{8}} = \left(\sum_{k=1}^{\frac{HW}{64}} \mathcal{A}_{i \times \frac{W}{8} + j,k} \mathcal{X}_{i \times \frac{W}{8} + j,k}, \sum_{k=1}^{\frac{HW}{64}} \mathcal{A}_{i \times \frac{W}{8} + j,k} \mathcal{Y}_{i \times \frac{W}{8} + j,k}\right), \tag{3}$$

where $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{\frac{HW}{64} \times \frac{HW}{64}}$ is the x and y coordinate of the 2D position matrix.

Fine-scale Refinement Since each pixel's optical flow at a coarse scale corresponds to four pixels at a finer scale, replicating the value along both the x-axis and y-axis is a reasonable approach for achieving upsampling. Then we map each pixel $\mathbf{x}$ of $\mathcal{E}_s^{\frac{1}{4}}$ to the predicted position $\mathbf{x}'$ in $\mathcal{E}_t^{\frac{1}{4}}$. Due to errors in the coarse-scale optical flow results, we define a local grid around $\mathbf{x}'$ for adjustment:

$$\mathcal{G}(\mathbf{x}') = \left\{\mathbf{x}' + \mathbf{d} | \mathbf{d} \in \mathbb{Z}^2, \|\mathbf{d}\|_1 \leq r\right\}, \tag{4}$$

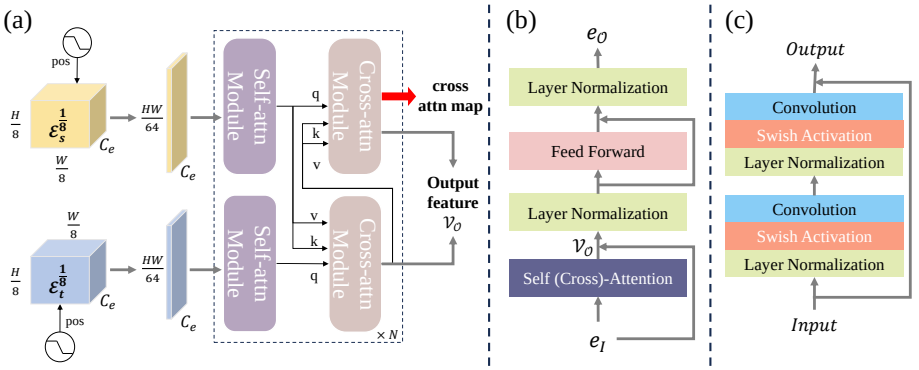

Fig. 2. The detailed network structures in the *Registration Module* and *Fusion Module*, comprising (a) the transformer architecture with a multi-head attention mechanism, (b) the layers of the Self-attention Module and Cross-attention Module, and (c) the specific network architecture of ResBlock.

which consists of integer offsets with a radius not exceeding r. Afterwards, based on $\mathcal{G}(\mathbf{x}')$, we employ bilinear sampling to acquire the neighboring features of $\mathbf{x}'$, which are then utilized to calculate the correlation with $\mathbf{x}$:

$$\mathcal{C}(\mathbf{x}) = \text{Concat}\left(\mathcal{E}_t^{\frac{1}{4}}(\mathbf{p}_1)\cdot\mathcal{E}_s^{\frac{1}{4}}(\mathbf{x}), \mathcal{E}_t^{\frac{1}{4}}(\mathbf{p}_2)\cdot\mathcal{E}_s^{\frac{1}{4}}(\mathbf{x}), \cdots, \mathcal{E}_t^{\frac{1}{4}}(\mathbf{p}_n)\cdot\mathcal{E}_s^{\frac{1}{4}}(\mathbf{x})\right) \tag{5}$$

where $\mathbf{p}_n \in \mathcal{G}(\mathbf{x}')$ and $\cdot$ denotes the matrix multiplication. Coarse-scale optical flow predictions help focus on the neighborhood, effectively reducing computational complexity and identifying potential accurate matches. Upsampled $\mathcal{F}_{up}^{\frac{1}{4}}(\mathbf{x})$ and $\mathcal{C}(\mathbf{x})$ are passed through two separate convolutional layers to obtain deep features, which are then concatenated and fused through a convolutional layer. The fused feature is further combined with $\mathcal{F}_{up}^{\frac{1}{4}}$ and the original image feature $\mathcal{E}_s^{\frac{1}{4}}$ to generate the optical flow adjustment $\Delta\mathcal{F}_{up}^{\frac{1}{4}}$. Finally, the refined optical flow $\mathcal{F}^{\frac{1}{4}}$ is calculated as $\mathcal{F}_{up}^{\frac{1}{4}} + \Delta\mathcal{F}_{up}^{\frac{1}{4}}$. By following the same procedure, we can derive the refined optical flow $\mathcal{F}_*$ at the original image resolution.

2.3 Fusion Module

The decoder's multi-scale structure, combined with its image generation capability, enables effective deep fusion of texture information. Therefore, we fuse the registered multi-scale deep features from two encoders with the corresponding scale features of the decoder using skip connections. This strategy allows us to retrieve and fuse surrounding information even in the presence of small registration errors, thereby reducing noise and restoring high-fidelity textures.

The predicted $\mathcal{F}_*$ is downsampled to construct an optical flow pyramid $\mathcal{F}_*^m(m = 1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8})$. This is combined with the multi-scale features of the encoder $\mathcal{E}_t^m$ and $\mathcal{E}_s^m$, as well as the multi-scale features of the source image decoder $\mathcal{D}_s^m$, to calculate the fusion feature $\mathcal{D}_{fuse}^m$. The process is defined as:

$$\begin{aligned}
\mathcal{E}^m_{reg} &= warp(\mathcal{F}^m_*, \mathcal{E}^m_t) \\
\mathcal{E}^m_{temp} &= \text{ResBlock}(\text{Concat}(\mathcal{E}^m_{reg}, \mathcal{E}^m_s)) \\
\mathcal{D}^m_{fuse} &= \text{ResBlock}(\text{Concat}(\mathcal{E}^m_{temp}, \mathcal{D}^m_s)).
\end{aligned} \tag{6}$$

where m means the feature scale, *warp* operator represents warping the target feature $\mathcal{E}^m_t$ using the optical flow $\mathcal{F}^m_*$ to obtain the registered feature $\mathcal{E}^m_{reg}$, and the detailed architecture of ResBlock is in Fig.2(c). Finally, the fused feature $\mathcal{D}^1_{fuse}$ is passed through the Layer Normalization, swish activation function, and a convolutional layer to generate the final texture map I_{fuse}.

2.4 Training Loss

The overall self-supervised objectives L_{self} is divided into three parts: Appearance loss L_{ap} consists of the L1 loss L_{rec} and SSIM loss L_{SSIM}, which are used to measure the pixel-wise difference and consistency between the predicted and GT texture map. Total variation loss L_{TV} is used to regularize the predicted optical flow $\mathcal{F}_p$ to be smooth. The perceptual loss L_p makes results I_{fuse} visually similar to input I_s following VQ-GAN[3]. The final loss is defined as:

$$L_{self} = \sum_{i=1}^{m} (\lambda_1 L_{ap}(I^m_{s'}, I^m_s) + \lambda_2 L_{TV}(\mathcal{F}^m_*)) + \lambda_1 L_{ap}(I_{fuse}, I_{gt}) + L_p(I_{fuse}, I_{gt}), \tag{7}$$

where m means the feature scale, λ_1, λ_2 are the weights of different loss terms, $I^m_{s'}$ and I^m_s represent the input and reconstructed source texture map using predicted optical flow $\mathcal{F}^m_*$ at scale m, and I_{fuse}, I_{gt} are the fused texture map and ground truth texture map.

3 Experiments

3.1 Datasets

To verify the effectiveness of the proposed framework, we use two public *in-vivo* datasets *SCARED*[1] and *Hamlyn*[17]. Drawing inspiration from [18], we randomly selected pairs of high-definition images with frame intervals of 1-10 frames from the original video and performed degradation operations on them to generate low-definition images, thus creating the training and test sets. After training on SCARED, we directly validated it across four test sets of the Hamlyn.

3.2 Experimental Settings

The experiments were conducted using the Pytorch and Lightning library on NVIDIA GeForce RTX 3090 GPUs. λ_1, λ_2 were set to 1, 0.1, respectively. The network was trained for 20 epochs with a batch size of 8, input/output resolution of 256x256, and utilized the Adam optimizer with a learning rate of 2.5e-5..

Table 1. Quantitative Results on Hamlyn Dataset. Higher values indicate better performance for all metrics. EndoMotion is abbreviated as E-M.

Methods	Video1(1305 frames)			Video2(1058 frames)		
	PSNR	SSIM	VIF	PSNR	SSIM	VIF
E-M	80.69±3.833	0.531±0.113	0.439±0.12	75.328±3.357	0.411±0.158	0.28±0.124
E-M+MBB	83.798±2.689	0.702±0.046	0.538±0.059	79.382±2.598	0.667±0.066	0.44±0.084
E-M+ours	**88.546±2.109**	**0.857±0.018**	**0.758±0.022**	**85.425±1.651**	**0.88±0.029**	**0.719±0.03**
	Video3(1529 frames)			Video4(342 frames)		
	PSNR	SSIM	VIF	PSNR	SSIM	VIF
E-M	82.236±3.447	0.492±0.132	0.418±0.122	74.919±2.442	0.463±0.129	0.301±0.104
E-M+MBB	84.781±2.104	0.674±0.05	0.509±0.058	78.463±1.754	0.692±0.051	0.446±0.067
E-M+ours	**89.578±2.028**	**0.845±0.019**	**0.746±0.035**	**84.451±1.001**	**0.884±0.023**	**0.7±0.025**

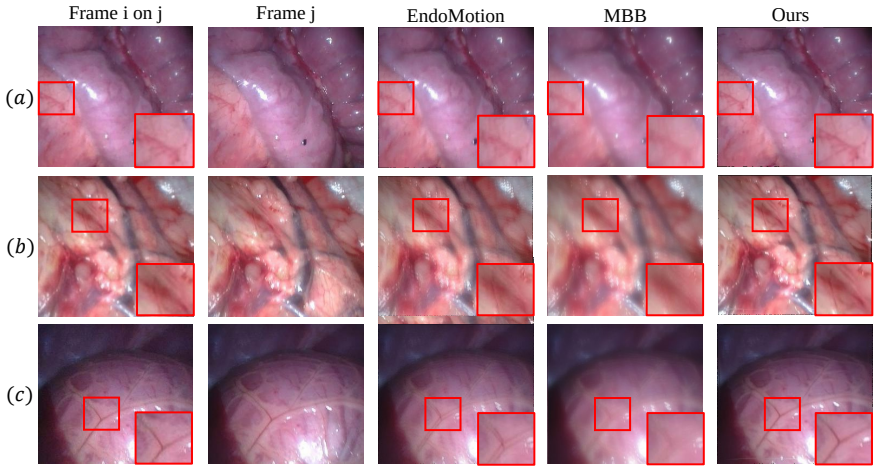

Fig. 3. Qualitative results of 2D texture consistency and quality on Hamlyn dataset. "Frame i on j" refers to the reprojection of frame i onto frame j. "EndoMotion" means the weighted sum of frame j and the reprojected frame i. "MBB" denotes the outcome of multi-band blending. "Ours" indicates the aligned texture using the predicted $\mathcal{F}_*$.

Since there are no ground truth for corresponding points, we adopted the Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR) and Visual information fidelity (VIF) as three criteria to assess the texture consistency and quality between adjacent frames.

3.3 Results and Analysis

We introduce EndoMotion [9] and EMDQ-SLAM [20] as comparisons, each representing a distinct approach for handling textures. We began by employing EndoMotion[1] to extract the depth and pose for each frame. Subsequently, leveraging the poses and 2D mapping detailed in Section 2.1, we projected the point cloud from the preceding frame onto the current one. With the captured image in

[1] https://github.com/UZ-SLAMLab/Endo-Depth-and-Motion

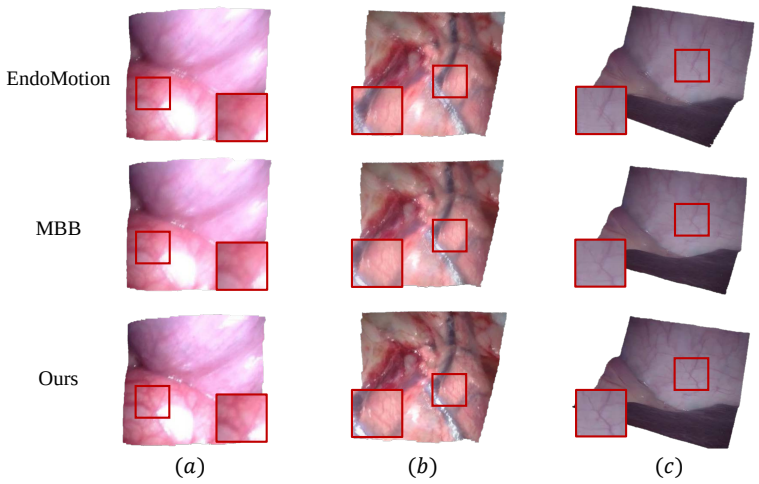

Fig. 4. 3D texture consistency and qualitative evaluation of two-frame fused mesh.

current frame, texture consistency can be assessed by computing the similarity. Given that EMDQ is not open-sourced, MBB[2] was employed for color blending between two frames to avoid texture artifacts. Our method fused the misaligned texture information from two frames, while providing the aligned texture map using predicted $\mathcal{F}_*$, which can be used for evaluation.

The quantitative results on the Hamlyn dataset are presented in Table 1. Any misalignment in features due to inaccurate optical flow $\mathcal{F}_*$ will result in noticeable ghosting or artifacts in images generated by the frozen VQ-GAN, making the textures even worse. Our proposed method outperforms the other two methods in all metrics, demonstrating its ability to generate textures with superior consistency and quality. Visual results are provided to illustrate the performance of different methods on both 2D images and 3D meshes, as shown in Fig.3 and Fig.4. EndoMotion exhibits misalignment blur in textures, while MBB, although mitigating this phenomenon, still reduces the clarity of textures. From the amplified details in rectangles, it can be observed that our method effectively eliminates the noise from the projected images by incorporating the relevant information from the current frame. Finally, our method effectively corrects texture misalignment, preserves the original texture details, and achieves a superior texture for the reconstructed mesh compared to the other two methods.

4 Conclusion

A method for in-vivo 3D texture optimization is proposed to address texture blurring or loss caused by inaccurate depth and pose estimation in the 3D reconstruction process. By utilizing camera projection and interpolation regression,

[2] https://github.com/CorentinBrtx/image-stitching

point cloud textures are transferred to a more manageable 2D space. An effective registration module, designed with features from a pre-trained generative network, aligns texture information from misaligned frames in a coarse-to-fine fashion. Subsequently, a hierarchical decoding architecture efficiently fuses information from two frames to eliminate noise. Extensive experiments on public datasets demonstrate the effectiveness and generalization across various laparoscopes. When pose estimation is inaccurate, the proposed method could select neighboring keyframes to enhance the 3D texture. Future research aims to extend the algorithm to entire video sequences, achieving incremental texture optimization within the SLAM framework.

Acknowledgments. This work is fully supported by the National Natural Science Foundation of China (62133004, 72293585 and 72188101), the Anhui Provincial Natural Science Foundation (2108085J33), and the Anhui Provincial Major Science and Technology Project (202203a05020010), the Fundamental Research Funds for the Central Universities (JZ2022HGPA0311, JZ2023HGQA0125 and JZ2024HGTA0174).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Allan, M., Mcleod, J., Wang, C., Rosenthal, J.C., Hu, Z., Gard, N., Eisert, P., Fu, K.X., Zeffiro, T., Xia, W., et al.: Stereo correspondence and reconstruction of endoscopic data challenge. arXiv preprint arXiv:2101.01133 (2021)
2. Eck, U., Wechner, M., Pankratz, F., Yu, K., Lazarovici, M., Navab, N.: Real-time 3d reconstruction pipeline for room-scale, immersive, medical teleconsultation. Applied Sciences **13**(18), 10199 (2023)
3. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
4. Haouchine, N., Dequidt, J., Peterlik, I., Kerrien, E., Berger, M.O., Cotin, S.: Image-guided simulation of heterogeneous tissue deformation for augmented reality during hepatic surgery. In: 2013 IEEE international symposium on mixed and augmented reality (ISMAR). pp. 199–208. IEEE (2013)
5. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (2023)
6. Li, L., Mazomenos, E., Chandler, J.H., Obstein, K.L., Valdastri, P., Stoyanov, D., Vasconcelos, F.: Robust endoscopic image mosaicking via fusion of multimodal estimation. Medical Image Analysis **84**, 102709 (2023)
7. Liu, X., Li, Z., Ishii, M., Hager, G.D., Taylor, R.H., Unberath, M.: Sage: slam with appearance and geometry prior for endoscopy. In: 2022 International conference on robotics and automation (ICRA). pp. 5587–5593. IEEE (2022)
8. Liu, Y., Li, C., Yang, C., Yuan, Y.: Endogaussian: Gaussian splatting for deformable surgical scene reconstruction. arXiv preprint arXiv:2401.12561 (2024)
9. Recasens, D., Lamarca, J., Fácil, J.M., Montiel, J., Civera, J.: Endo-depth-and-motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints. IEEE Robotics and Automation Letters **6**(4), 7225–7232 (2021)

10. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
11. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
12. Song, J., Wang, J., Zhao, L., Huang, S., Dissanayake, G.: Mis-slam: Real-time large-scale dense deformable slam system in minimal invasive surgery based on heterogeneous computing. IEEE Robotics and Automation Letters **3**(4), 4068–4075 (2018)
13. Song, J., Zhang, R., Zhu, Q., Lin, J., Ghaffari, M.: Bdis-slam: a lightweight cpu-based dense stereo slam for surgery. International Journal of Computer Assisted Radiology and Surgery pp. 1–10 (2024)
14. Sun, X., Wang, F., Ma, Z., Su, H.: Dynamic surface reconstruction in robot-assisted minimally invasive surgery based on neural radiance fields. International Journal of Computer Assisted Radiology and Surgery pp. 1–12 (2023)
15. Wang, Y., Long, Y., Fan, S.H., Dou, Q.: Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 431–441. Springer (2022)
16. Widya, A.R., Monno, Y., Imahori, K., Okutomi, M., Suzuki, S., Gotoda, T., Miki, K.: 3d reconstruction of whole stomach from endoscope video using structure-from-motion. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 3900–3904. IEEE (2019)
17. Ye, M., Johns, E., Handa, A., Zhang, L., Pratt, P., Yang, G.Z.: Self-supervised siamese learning on stereo image pairs for depth estimation in robotic surgery. arXiv preprint arXiv:1705.08260 (2017)
18. Zhang, K., Liang, J., Van Gool, L., Timofte, R.: Designing a practical degradation model for deep blind image super-resolution. In: IEEE International Conference on Computer Vision. pp. 4791–4800 (2021)
19. Zhang, X., Ji, X., Wang, J., Fan, Y., Tao, C.: Renal surface reconstruction and segmentation for image-guided surgical navigation of laparoscopic partial nephrectomy. Biomedical Engineering Letters pp. 1–10 (2023)
20. Zhou, H., Jayender, J.: Emdq-slam: Real-time high-resolution reconstruction of soft tissue surface from stereo laparoscopy videos. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24. pp. 331–340. Springer (2021)
21. Zhou, H., Jayender, J.: Real-time nonrigid mosaicking of laparoscopy images. IEEE transactions on medical imaging **40**(6), 1726–1736 (2021)