ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447800

SEAM MASK GUIDED PARTIAL RECONSTRUCTION WITH QUANTUM-INSPIRED LOCAL AGGREGATION FOR DEEP IMAGE STITCHING

Chen-Bin Feng[†], *Jie Zhang*[†], *Jiaxue Li*, *Yicong Zhou*[‡]

Department of Computer and Information Science, University of Macau, Macau, China.

ABSTRACT

In image stitching, artifacts caused by misalignment affect the visual quality and the performance of subsequent tasks such as segmentation and detection. This paper proposes SMPR, a reconstruction-based aligned image composition method to minimize artifacts. SMPR fuses images in part of the overlapping areas and reconstructs other portions from single images. Specifically, we propose a seam mask generation method to obtain optimal seam masks that pass through minimal misalignment. During training, we use the seam masks to guide the model in detecting optimal fusion areas. In testing, the model can detect fusion areas without seam masks and reconstruct stitching results. We propose a quantum-inspired local aggregation (QILA) module to improve feature reconstruction performance. We develop an encoder-decoder network with QILA and experiment on a real-world dataset. The experiments show that our method outperforms state-of-the-art methods in both qualitative and quantitative aspects.

Index Terms— Image Stitching, Image Reconstruction, Deep Learning, Image Processing, Quantum Neural Network.

1. INTRODUCTION

Image stitching is a meaningful task in multimedia signal processing. The goal of image stitching is to create large-view images by combining images taken from different perspectives. It has a lot of applications, including surveillance video [1], autonomous driving [2], remote sensing [3], and UAV (Unmanned Aerial Vehicle) imaging[4].

Most image stitching methods can be divided into two steps: image alignment and image composition. Image alignment converts images captured from different positions to the same plane and aligns as much content as possible. Due to parallax between input images, it is difficult for alignment methods to align all the image content well. So, the stitching results usually suffer from artifacts if we directly fuse aligned images. Therefore, we need image composition methods to produce better stitching results. The commonly used image composition technique is seam detection. Two-image seam detection methods detect the optimal seam that passes through regions with minimum misalignment on the overlapping area between two images. The seam separates the output image into two parts: one is the image content from the first image, and the other is from the second image. These two parts compose the stitching result. After seam detection and composition, we use image blending methods such as Poisson image blending [5] to smooth the seam for a natural transition. However, these image blending methods cannot eliminate large misalignments and may produce artifacts around seams.

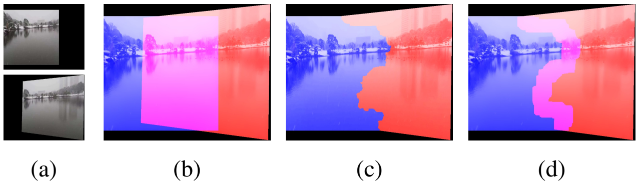


Fig. 1: Three kinds of image composition methods for image stitching. (a) Input images. (b) Original reconstruction. (c) Seam detection. (d) Our partial reconstruction. The pink area represents the fusion area. The blue and red regions are content from a single input image.

Recently, some deep-learning-based methods [6, 7] use feature reconstruction with perceptual loss [8] to reduce misalignment. The perceptual loss is calculated on the feature level instead of the pixel level. Thus, it can reconstruct image content and reduce artifacts. However, when misalignment is large, the reconstruction results may have artifacts. Our method aims to reduce artifacts in large parallax cases via partial reconstruction. Different from original reconstruction methods that fuse whole overlapping areas, we use seam masks to guide the model to fuse smaller areas. This is because smaller areas have less misalignment. We hope to use smaller fusion areas to reduce possible artifacts in the stitching results. The original reconstruction strategy is illustrated in (b) of Figure 1, the seam detection method is illustrated in (c), and our partial reconstruction is illustrated in (d). Compared to seam detection methods, our method dilates the seam line to make it a banded area and uses it for two-image feature reconstruction.

*: This work was supported by the Science and Technology Development Fund, Macau SAR (File no. 0049/2022/A1), and by the University of Macau (File no. MYRG2022-00072-FST).

†: Equal Contribution

‡: Corresponding Author: Yicong Zhou (yicongzhou@um.edu.mo)

Inspired by the extraordinary feature mixing performance of quantum networks, we work on developing quantum network architecture for reconstruction. Wave MLP [9] is a quantum network on the patch level that applies to image classification, segmentation, and object detection. For our task, we focus on local feature mixing. So, we aim to propose a quantum network on the pixel level to improve local feature aggregation performance for reconstruction.

In this paper, we propose a reconstruction-based image composition method, SMPR. It transforms aligned inputs to stitching results with fewer artifacts and better visual quality. The contributions of this paper are summarized as follows:

- We propose a pixel-level quantum-inspired reconstruction network. To the best of our knowledge, it is the first quantum network for deep image stitching. We verify that it can improve the reconstruction performance.
- We propose a seam mask guided partial reconstruction method using generated seam masks. It can reduce artifacts related to misalignment.
- Extensive experiments show that our method outperforms existing methods regarding the number of artifacts and the visual quality of stitching results.

2. METHODOLOGY

Our methodologies include quantum-inspired local aggregation, seam mask generation, and seam mask-guided partial reconstruction.

2.1. Quantum-Inspired Local Aggregation

For the aligned image reconstruction task, we need to aggregate local information for better feature mixing. We propose a local information aggregation module under quantum representations. Concretely, we represent each pixel as a wave that contains both amplitude and phase. We convert real domain pixels to the complex domain:

$$\tilde{z_j} = |z_j| \odot e^{i\theta_j}, j = 1, 2, 3, \cdots, H \times W \times C, \quad (1)$$

where $\tilde{z_j}$ represents the pixels in complex domain, $|z_j|$ stands for the amplitude, θ_j denotes phase, i stands for the imaginary unit, $\odot$ is the pixel-level multiplication, H, W stand for height and width, C is channel number. However, it is hard to develop neural network architectures using complex domain representation. So we use Euler's formula to represent each pixel in real part and imaginary part:

$$\tilde{z_j} = |z_j| \odot cos\theta_j + i|z_j|sin\theta_j, j = 1, 2, 3, \cdots, H \times W \times C, \quad (2)$$

We use equation (2) as the quantum representation of wave-like pixels.

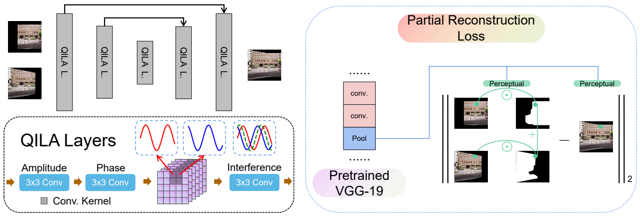


Fig. 2: The reconstruction networks, the detail of QILA layers, and the illustration of partial reconstruction loss.

Wave interference is a physical phenomenon of two waves meeting when traveling in the same medium. We use convolutional neural networks (CNN) to represent wave interference. The convolution operation can be seen as multi-pixel wave interference. With CNN, each pixel has a large receptive field to fuse more local features.

In detail, we need pixel-level amplitude $|z_j|$ and phase θ_j to do local aggregation. Given an input feature map F $\in \mathbb{R}^{H \times W \times C1}$, we use a 2D convolutional layer and ReLU activation to get A $\in \mathbb{R}^{H \times W \times C2}$, which works as the amplitude. Then we input A to another 2D convolutional layer and ReLU activation to get P $\in \mathbb{R}^{H \times W \times C3}$, which works as the phase. Finally, we get the quantum representation, which is the channel dimension concatenation of $A \odot cos(P)$ and $A \odot sin(P)$. Then, we use the following CNN layers to do the pixel-level wave interference. The illustration of the QILA layers can be found in Figure 2. Compared to normal CNN, quantum-inspired local aggregation can fully mix local information and generate better reconstruction results.

2.2. Seam Mask Generation

We use input images aligned by an existing homography estimation method [10]. Then, we use the generated seams to produce seam masks. We expand the seam areas so the model can fuse the misalignments around the seam areas of two images through feature reconstruction. The mask generation process is illustrated in Figure 3. The specific steps are as follows,

$$S_{line}, S_{M_1}, S_{M_2} = SD(I_1, I_2), \quad (3)$$

where the SD is the seam detection algorithm, we use Graph Cut [11]. The inputs are two aligned images. S_{line} is the seam line that the width is 1. S_{M_1} and S_{M_2} are the two seam masks respectively. Then we expand the S_{line} using morphological dilation operation [12]. The M_{SL} is the output seam band. The width of M_{SL} is 30 pixels here. We notice that the seam band exceeds the boundary of the input image. So, we use the overlapping mask to remove the redundant parts.

$$M_{overlapping} = M_A \cap M_B, \quad (4)$$

$$M'_{SL} = M_{SL} \cap M_{overlapping}, \quad (5)$$

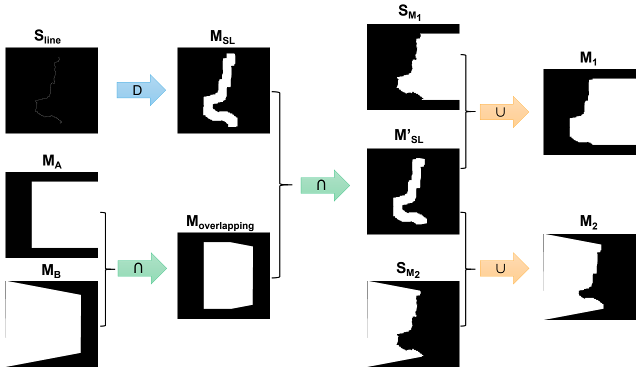


Fig. 3: The process of seam mask generation. D stands for morphological dilation.

where M_A and M_B are masks that stand for the valid areas of inputs, and $\cap$ denotes the intersection of two images. Then, we use this seam band to produce the final masks:

$$M_1 = S_{M_1} \cup M'_{SL}, \tag{6}$$

$$M_2 = S_{M_2} \cup M'_{SL}, \tag{7}$$

where $\cup$ denotes the union of two images. The overlapping area of the two final masks is the seam band area. It is the image fusion area.

2.3. Seam-Mask-Guided Partial Reconstruction

We use the aligned input images and the generated seam masks for partial reconstruction. We use the seam masks for training and do not use them during testing. The model learns to detect fusion areas and reconstruct fusion content during testing.

2.3.1. *Reconstruction Network*

We propose a U-shaped network with QILA layers as illustrated in the left top corner of Figure 2. We use QILA layers as the basic feature extraction blocks of the encoders and decoders. We concatenate the features from the downsampling stage to the upsampling stage to restore image details.

2.3.2. *Partial Reconstruction Loss*

SMPR is trained with pixel-level perceptual loss as shown in Figure 2. At the same time, it can learn to detect fusion areas with given seam masks. Concretely,

$$\begin{aligned}\mathcal{L}_{PR} = &\|\Phi(I_1 \odot M_1) - \Phi(O \odot M_1)\|_2 \\ &+ \|\Phi(I_2 \odot M_2) - \Phi(O \odot M_2)\|_2\,,\end{aligned} \tag{8}$$

where $\mathcal{L}_{PR}$ stands for partial reconstruction loss, Φ means the output feature maps of pretrained VGG-19 [13], I_1 and I_2 stand for input 1 and input 2, M_1 and M_2 stand for mask 1 and mask 2, O denotes the output.

3. EXPERIMENT

3.1. Dataset

We use a real-world dataset UDIS-D [7] for training and testing. The UDIS-D contains 10,440 training image pairs and 1,106 testing image pairs. We train the model on the aligned 10,440 training image pairs and the proposed 10,440 seam mask pairs. We test the model on the UDIS-D testing dataset.

3.2. Metrics

The dataset does not have ground truth so we use a combination of manual and algorithmic measurements. The manual metric is the failure cases and success rates. The failure cases refer to stitching results with unacceptable artifacts. The algorithmic metrics are the non-reference image quality assessment methods. We use BRISQUE [14] and PIQE [15] as our image quality assessment methods.

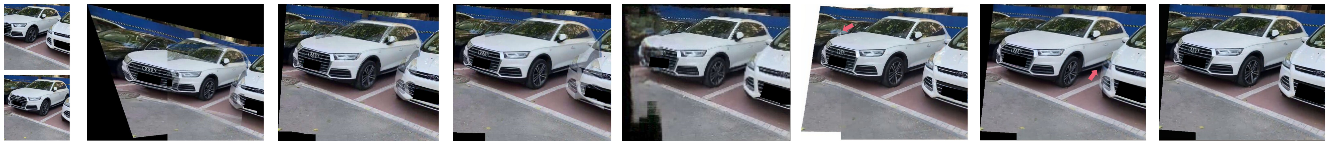
3.3. Performance Comparison

3.3.1. *Comparisons of Complete Image Stitching*

We input non-aligned image pairs and evaluate the performance of different complete image stitching methods. We train our model on aligned UDIS-D and seam masks for five epochs. To evaluate the performance of different scenes, we split the 1,106 testing image pairs into three categories, i.e., large parallax (78 image pairs), small parallax (923 image pairs), and low texture (105 image pairs). We compare with image stitching methods, including traditional methods SIFT [16] combined with RANSAC [17], APAP [18], robust ELA [19], and deep learning methods VFISNet [6], EPIS [20], and UDIS [7]. The failure cases, success rate, and image quality assessment score are illustrated in Table 1. We can see that traditional methods (lines one to three) perform badly in low-texture cases. Deep learning methods (lines four to six) perform badly in large parallax cases. Our method outperforms other methods in large parallax, small parallax, and low texture, obtaining the fewest failure cases and the highest success rate. For the image quality assessment metrics, our method obtains first place, and the traditional method SIFT gets second place. The APAP and robust ELA get F (Failure) in IQA. This is because they have failed stitching results and cannot be calculated. From the qualitative comparison results illustrated in Figure 4, we notice that SIFT, APAP, robust ELA, VFISNet, and UDIS have artifacts in the content, while EPIS has distortion at the intersections and edges. The result of our SMPR does not have these problems.

3.3.2. *Comparisons of Image Composition*

We use aligned image pairs as inputs to compare the performance with three seam detection methods, including dynamic programming (DP) [21], Voronoi [22], and Graph Cut (GC)



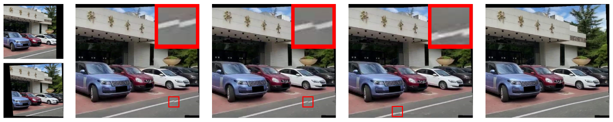
(a) (b) SIFT (c) APAP (d) Rob. ELA (e) VFISNet (f) EPIS (g) UDIS (h) SMPR

Fig. 4: Qualitative comparison of different complete image stitching methods on the non-aligned image inputs (a).

Method	Large Parallax		Small Parallax		Low Texture		Overall		IQA	
	Failure	Rate	Failure	Rate	Failure	Rate	Failure	Rate	BRI.(↓)	PIQE(↓)
SIFT [16]	26	66.67%	43	95.34%	48	54.29%	117	89.42%	36.387	20.403
APAP [18]	23	70.51%	12	98.70%	25	76.19%	60	94.58%	F	F
Rob. ELA [19]	39	50.00%	44	95.23%	67	36.19%	150	86.44%	F	F
VFISNet [6]	31	60.26%	100	89.17%	30	71.43%	161	85.44%	48.275	22.523
EPIS [20]	25	67.95%	23	97.51%	6	94.29%	54	95.12%	55.846	20.099
UDIS [7]	32	58.97%	1	99.89%	**3**	**97.14%**	36	96.75%	40.155	20.559
SMPR (Ours)	**20**	**74.36%**	**0**	**100.00%**	**3**	**97.14%**	**23**	**97.92%**	**35.350**	**19.874**

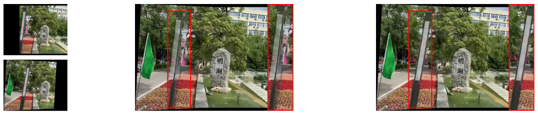
Table 1: Success rate and image quality assessment metrics of different complete image stitching methods on different scenes of 1106 non-aligned UDIS-D testing images.

[11]. A qualitative comparison is illustrated in Figure 5. From the figure, we notice that all the seam detection methods have artifacts. These artifacts are around the seams. Our partial reconstruction does not have such problems because we detect fusion areas and reconstruct features.



(a) (b) DP (c) Voronoi (d) GC (e) SMPR

Fig. 5: Qualitative comparison of different image composition methods on the aligned image inputs (a).



(a) (b) w/o SMG (c) w/ SMG

Fig. 6: Ablation study of result without seam mask guidance (SMG) (b) and with SMG (c) on the aligned image inputs (a).

3.4. Ablation Study

We perform ablation studies on seam mask guidance (SMG) and QILA layers. The model without QILA retains the same number of CNN layers. From Table 2, we know that the model with SMG gets fewer failure cases than models without SMG. From Table 3, we can see that the model with QILA layers gets better image quality scores. From these results, we conclude that (1) QILA helps improve reconstruction quality. (2) SMG decreases failure cases. From Figure 6, we can see that the model without SMG cannot deal with misalignment and produce artifacts, while the model with SMG decreases artifacts via fusion area detection and feature reconstruction.

Method	Failure	Success Rate
w/o SMG	35	96.84%
Ours	23	97.92%

Table 2: Ablation study of SMG on 1106 images of UDIS-D.

Method	BRISQUE(↓)	PIQE(↓)
w/o QILA	36.089	20.411
Ours	35.350	19.874

Table 3: Ablation study of QILA on 1106 images of UDIS-D.

4. CONCLUSION

In this paper, we proposed an alignment image composition method, SMPR. In particular, the seam mask generation method produces seam masks. The seam mask guided partial reconstruction model performs fusion area detection and visual reconstruction. The quantum-inspired local aggregation (QILA) better mixes features and produces stitching results with better visual quality. It has been demonstrated in qualitative and quantitative experiments that our SMPR outperforms other image stitching methods.

5. REFERENCES

[1] Anqi Zhu, Lin Zhang, Juntao Chen, and Yicong Zhou, "Pedestrian-aware panoramic video stitching based on a structured camera array," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 17, no. 4, pp. 1–24, 2021.

[2] Mostafa Rizk, Ahmad Mroue, Mohammad Farran, and Jamal Charara, "Real-time slam based on image stitching for autonomous navigation of uavs in gnss-denied regions," in *2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*. IEEE, 2020, pp. 301–304.

[3] Yujie Zhang, Zhiying Wan, Xingyu Jiang, and Xiaoguang Mei, "Automatic stitching for hyperspectral images using robust feature matching and elastic warp," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 3145–3154, 2020.

[4] Kuo-Liang Chung and Dai-Yu Row, "An adaptive joint bilateral interpolation-based color blending method for stitched uav images," *Remote Sensing*, vol. 14, no. 21, pp. 5440, 2022.

[5] Patrick Pérez, Michel Gangnet, and Andrew Blake, "Poisson image editing," in *ACM SIGGRAPH 2003 Papers*, pp. 313–318. 2003.

[6] Lang Nie, Chunyu Lin, Kang Liao, Meiqin Liu, and Yao Zhao, "A view-free image stitching network based on global homography," *Journal of Visual Communication and Image Representation*, vol. 73, pp. 102950, 2020.

[7] Lang Nie, Chunyu Lin, Kang Liao, Shuaicheng Liu, and Yao Zhao, "Unsupervised deep image stitching: Reconstructing stitched features to images," *IEEE Transactions on Image Processing*, vol. 30, pp. 6184–6197, 2021.

[8] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 694–711.

[9] Yehui Tang, Kai Han, Jianyuan Guo, Chang Xu, Yanxi Li, Chao Xu, and Yunhe Wang, "An image patch is a wave: Phase-aware vision mlp," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10935–10944.

[10] Liu Kai and Ge Teng, "Unsupdis-pytorch," Available: https://github.com/liudakai2/UnsupDIS-pytorch, 2022.

[11] Vivek Kwatra, Arno Schödl, Irfan Essa, Greg Turk, and Aaron Bobick, "Graphcut textures: Image and video synthesis using graph cuts," *Acm transactions on graphics (tog)*, vol. 22, no. 3, pp. 277–286, 2003.

[12] Robert M Haralick, Stanley R Sternberg, and Xinhua Zhuang, "Image analysis using mathematical morphology," *IEEE transactions on pattern analysis and machine intelligence*, vol. PAMI-9, no. 4, pp. 532–550, 1987.

[13] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[14] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on image processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[15] N Venkatanath, D Praneeth, Maruthi Chandrasekhar Bh, Sumohana S Channappayya, and Swarup S Medasani, "Blind image quality evaluation using perception based features," in *2015 twenty first national conference on communications (NCC)*. IEEE, 2015, pp. 1–6.

[16] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, pp. 91–110, 2004.

[17] Martin A Fischler and Robert C Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[18] Julio Zaragoza, Tat-Jun Chin, Michael S Brown, and David Suter, "As-projective-as-possible image stitching with moving dlt," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2339–2346.

[19] Jing Li, Zhengming Wang, Shiming Lai, Yongping Zhai, and Maojun Zhang, "Parallax-tolerant image stitching based on robust elastic warping," *IEEE Transactions on multimedia*, vol. 20, no. 7, pp. 1672–1687, 2017.

[20] Lang Nie, Chunyu Lin, Kang Liao, and Yao Zhao, "Learning edge-preserved image stitching from multi-scale deep homography," *Neurocomputing*, vol. 491, pp. 533–543, 2022.

[21] Marie-Lise Duplaquet, "Building large image mosaics with invisible seam lines," in *Visual information processing VII*. SPIE, 1998, vol. 3387, pp. 369–377.

[22] Franz Aurenhammer and Rolf Klein, "Voronoi diagrams.," *Handbook of computational geometry*, vol. 5, no. 10, pp. 201–290, 2000.