Complex & Intelligent Systems (2025) 11:90
https://doi.org/10.1007/s40747-024-01702-x

ORIGINAL ARTICLE

Progressive alignment and interwoven composition network for image stitching

Xiaoting Fan[1] · Long Sun[2] · Zhong Zhang[1] · Tariq S. Durrani[3]

Received: 2 June 2024 / Accepted: 7 September 2024

Abstract

As one of the fundamental tasks in computer graphics and image processing, image stitching aims to combine multiple images with overlapping regions to generate a high-quality naturalness panorama. Most deep learning based image stitching methods suffer from unsatisfactory performance, because they neglect the cooperation relationship and complementary information between reference image and target image. To address these issues, we propose a progressive alignment and interwoven composition network (PAIC-Net) to produce satisfactory panorama images, which learns the cooperation relationship by a progressive homography alignment module and captures the complementary information by an interwoven image composition module. Specifically, a progressive homography alignment module is presented to align the input images, which progressively warps the reference and target images by focusing more on the combination of self-features and cooperation features. Then, an interwoven image composition module is presented to seamlessly fuse aligned image pairs, where the complementary information of one-view is captured to guide another-view in an interweaved way. Finally, an alignment loss and a composition loss are introduced to reduce alignment distortions and enhance seam consistency of the final image stitching results. Experimental results on benchmark datasets demonstrate that PAIC-Net outperforms state-of-the-art image stitching methods both quantitatively and qualitatively.

Keywords Image stitching · Deep learning · Progressive homography Alignment · Interwoven image composition

Introduction

Image stitching is an extremely hot topic in the field of multimedia display technology and computer graphics, which aims to generate a high-quality wide field panorama images from multiple images. This technology has played an essential role in various applications nowadays, such as immersive communication [1, 2], remote sensing [3], virtual reality and augmented reality [4]. However, the inconsistencies of the overlapping regions between reference and target images may cause obvious alignment distortions and seam artifacts. Therefore, how to achieve natural panorama image from wide field is a challenging task.

At present, traditional image stitching methods include global homography methods and local warping methods [5]. Global homography methods tend to match complicated geometric features and estimate global homography relationship, such as Autostitch [6] and dual-homography [7]. However, when the image scene is not coplanar or contains more than one depth, these global homography methods may introduce serious structural distortions in the overlapping regions. To improve stitching performance for the cases with parallax, local warping methods divide the image pairs into uniform cells and construct local parametric warping constraints, including as-projective-as-possible (APAP) warps [8] and robust elastic warping [9]. These methods focus on

✉ Zhong Zhang
zhong.zhang8848@gmail.com

Xiaoting Fan
xtfan@tjnu.edu.cn

Long Sun
sunlong@tjcu.edu.cn

Tariq S. Durrani
t.durrani@strath.ac.uk

[1] Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin 300387, China

[2] School of Information and Engineering, Tianjin University of Commerce, Tianjin 300134, China

[3] Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, Scotland, UK

Published online: 23 December 2024

reducing the incorrect alignment in the overlapping regions. Unfortunately, they are likely to cause the local inconsistent, such as stretched curves and non-uniformly planes.

Due to the strong representation ability and flexible structure of convolution neural network (CNN), some deep CNN-based image stitching methods [10–14] have achieved SOTA performance recently. In general, deep CNN-based image stitching methods utilize a deep homography to align images and then compose aligned images to produce panorama [15–17]. For instances, some CNN-based deep homography estimation networks [18, 19] were applied to warp reference and target images for image stitching, while in [20, 21], some CNN-based image composition networks were designed to fuse aligned image pairs to generate panorama by learning the edge information. These methods can effectively process the low-texture scenes and unnatural cases. Nevertheless, the shortcomings of existing deep CNN-based image stitching methods are two folds: (1) For image alignment stage, some methods only utilize the information in reference image or target image itself and do not consider the cooperation relationship between reference image and target image, which introduces alignment distortions when handling the scenes with different parallax; (2) For image composition stage, existing methods usually adopt a simple deep CNN composition strategy to fuse aligned image pairs, which ignores the complementary information between aligned image pairs, resulting in unsatisfying panorama image with obvious seam artifacts.

To overcome the above-mentioned drawbacks, we present a progressive alignment and interwoven composition network (PAIC-Net) for image stitching, which can reduce alignment distortions and eliminate seam artifacts effectively. The major contributions of the proposed PAIC-Net are summarized as follows.

1. To align the input images and prevent alignment inconsistency, a progressive homography alignment sub-network is presented to progressively warp the reference and target images by leveraging the self-features of image itself and cooperation features between image pairs.
2. In order to fuse aligned image pairs and reduce seam artifacts, an interwoven image composition sub-network is proposed to integrate the complementary information between two aligned image pairs. The interwoven image composition sub-network is composed of four interwoven swin transformer modules in an interweaved way.
3. The proposed PAIC-Net significantly exceeds SOTA methods in two image stitching quality metrics on various datasets, which verifies the effectiveness of proposed PAIC-Net.

The organizational architecture of the paper is as follows. “Related work” section summarizes the related work of traditional image stitching methods, deep learning-based image stitching methods, and vision transformer. “Proposed method” section introduces a detailed discussion of the PAIC-Net. In “Experiments” section, qualitative and quantitative comparisons are presented, and the ablation studies are performed in details. The limitation and future work is introduced in “Limitations and future work” section and “Conclusion” section remarks the conclusions of the paper.

Related work

Traditional image stitching methods

Image stitching has been extremely popular in the past decades, with approaches that are divided into global homography methods and local warping methods. As a representative work, Autostitch technology [6] used feature matching, homography estimation, and multi-band blending to stitch multiple images, which enabled users to synthesize panorama images without any stitching foundation. In addition, Gao et al. [7] introduced a dual-homography estimation module to construct seamless image panoramas. However, global homography methods usually perform global geometric deformation on images, which may cause misalignment artifacts or local ghosting. To address these issues, some local warping methods have been designed to promote the image stitching performance. For instance, a smoothly varying affine stitching method [22] was presented to handle large parallax. Zaragoza et al. [8] designed a moving direct linear transformation framework to tweak the APAP warps. Inspired by the APAP method, Lin et al. [23] combined the local homography and global similarity transformations for warping and fusing image pairs.

In order to obtain good alignment and minimal local distortions simultaneously, Chen et al. [24] used a global similarity prior to constrain the similarity transformation. In [9], a robust elastic warping method was presented to generate natural-looking panoramas, where a global transformation was combined with global similarity transformation to mitigate projective distortions. Meanwhile, a manifold optimization energy function [25] was proposed to estimate spatially varying homographies between image pairs to realize image alignment. Additionally, Liao et al. [26] introduced a parametric warp module and a mesh warp module for natural image stitching. Similarly, Zhang et al. [27] constructed a layered warping constraint to stitch natural images with large parallax. To mitigate ghosting and preserve structure, Xue et al. [28] designed a linear structures to align images and a seam measurement to compose image. In addition, Zhang et al. [29] combined the rectangular boundaries to produce panorama images without content artifacts.

Deep image stitching methods

Recently, it has already been demonstrated that deep image stitching methods show great advantages compared with traditional methods. As the CNN has a good performance in feature extraction and matching, some deep homography estimation methods have been explored to align image pairs. These methods are applied to find the global perspective transformation between reference and target images. For example, Yan et al. [30] summarized multi-viewpoint image stitching methods based on deep learning, and introduced some evaluation metrics and experimental results to demonstrate the performance of different image stitching methods. In addition, Zhang et al. [31] introduced an unsupervised training way to estimate a deep homography, which extracted an outlier mask to choose reliable regions and constructed a triplet loss for estimating homography. These methods are applied to find the global perspective transformation between reference and target images.

Following the success of deep homography alignment, some deep image stitching methods are presented to generate wide field-of-view panorama image. In [18], Nie et al. presented a multi-grid homography estimation method, where a contextual correlation layer (CCL) was designed to utilize the feature correlation between reference and target images. In [20, 21], an edge-preserved deformation module and an edge-guided composition module were proposed to produce artifact-free stitching images, respectively. Considering the case of small parallax, a homography estimation network was designed to warp images and an image content loss function was designed to reduce shape distortions [32]. Inspired the idea of image semantic information under perspective geometry, Li et al. [10] designed some local transformation models to constrain matched regions, which promoted accurate image alignment. Different from supervised networks, an unsupervised deep image stitching method [11] was first introduced to combine images with few features or low resolution. Similarity, a parallax-tolerant unsupervised deep image stitching method [33] was designed to handle large-parallax cases, which performed a robust warp to model the image registration. In order to realize free-view image stitching, Xie et al. [34] designed a fast lightweight image reconstruction method, where a ShuffleNet was applied to extract feature maps and an optimized loss was utilized to reduce content distortions. In [35], a panoramic image stitching method based on deep CNN was presented and a novel panoramic image generation dataset was introduced to evaluate the image stitching performance.

Vision transformer

Recent years, transformer has been applied in the field of computer vision community successfully, because of the strong ability of learning the correlation between two pixels that are far apart. Numerous of vision transformer-based models have achieved competitive performance in numerous vision tasks, such as object detection, image segmentation, pose estimation, visual recognition, and classification [36, 37].

One of the early attempts of exploring transformer in the vision task is the object detection [38]. Similarity, Zheng et al. [39] designed an image semantic segmentation transformer module by treating the image segmentation as a prediction task. In addition, Gao et al. [40] designed a facial structure attention and a super-resolution transformer module to improve the structure restoration of face image. In [41], a contextual vision transformer network was presented for visual recognition. Furthermore, Patrick et al. [42] introduced a high-resolution image synthesis method based on vision transformer, where the transformers were used to model the context vocabulary in high-resolution image. Li et al. [43] designed a convolution based vision transformer method for infrared and visible image fusion.

Proposed method

Overview

Existing image stitching methods usually only explore the information of image pairs themselves for aligning and composing image pairs, which ignores the cooperation relationship and complementary information between image pairs, resulting in unsatisfying image stitching results. Therefore, instead of building an independent image stitching network for reference image and target image respectively, it is necessary to leverage the cooperation features and complementary features between image pairs to improve image stitching performance. Inspired by this, we present a Progressive Alignment and Interwoven Composition network (PAIC-Net) for image stitching, which includes progressive homography alignment, interwoven image composition, an alignment loss and a composition loss. Figure 1 shows the overview architecture of the PAIC-Net.

Given a reference image and a target image with overlapping regions, the proposed PAIC-Net can output a stitched panorama image with a wide-angle view. Firstly, a progressive homography alignment sub-network is presented to warp the overlapping regions between input image pairs and prevent alignment inconsistency, where a cross feature cooperation module (CFCM) is designed to obtain the refined cross features for estimating deep homography. Then, an interwoven image composition sub-network is designed to combine aligned images, which merges the complementary information between two aligned views by interwoven swin transformer module (ISTM). Finally, an alignment loss is

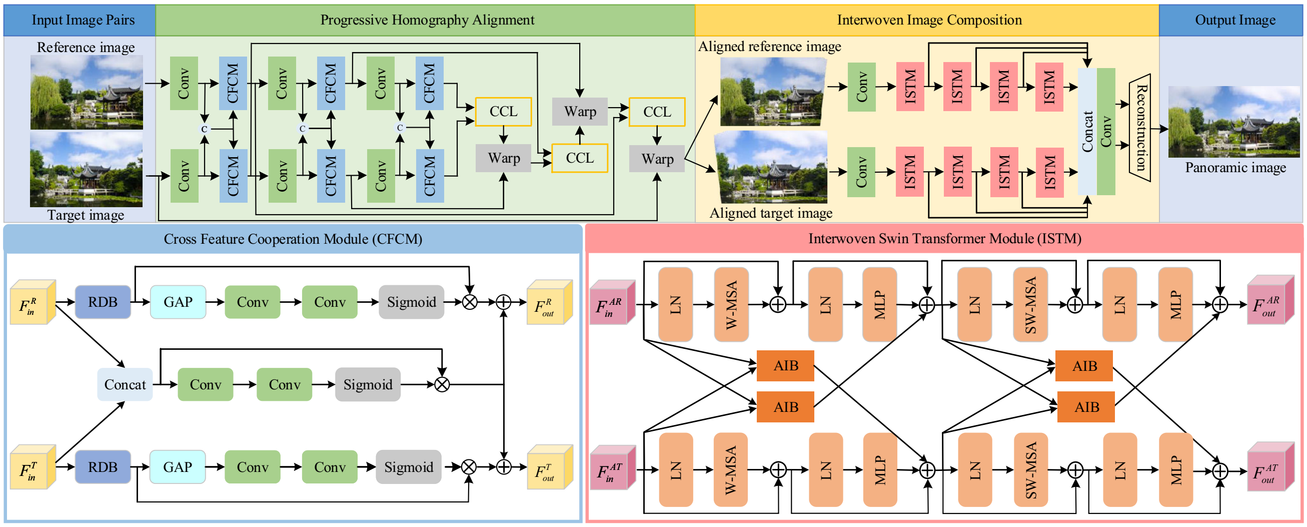

Fig. 1 Overview of the proposed PAIC-Net

employed to reduce the geometric distortions and a composition loss is applied to reduce seam artifacts. Next, we will provide a detailed analysis of the proposed PAIC-Net.

Progressive homography alignment

At present, how to align the overlapping regions between input images accurately by estimating a deep homography is one of the challenges for image stitching. Existing image stitching methods [12, 32] usually only use the information of image pairs themselves to obtain the deep features for estimating homography, which may lead to image misalignment. To filter out the misleading information when extracting multi-level cross features of reference and target images, we propose the progressive homography alignment sub-network to align image pairs progressively by considering that the cooperation relationship between reference and target images. As shown in Fig. 1, the progressive homography alignment sub-network includes the reference and target streams, each stream contains three stages, and each stage mainly consists of a convolutional layers, a CFCM, a contextual correlation layer (CCL) [18], and a spatial transformer module (Warp) [44]. For each stage in the progressive homography alignment sub-network, the CCL employs the refined cross features in the previous CFCM as a guidance to learn the multi-level cross information of reference and target images fed into the current stage.

As the key component of the proposed progressive homography alignment sub-network, the CFCM will be introduced in this section. Since the deep spatial and channel features of one view are usually somewhat essential for another view to learn useful information in estimating homography, a CFCM is designed to obtain refined cross features of reference and target images for homography estimation progressively. In each CFCM, the channel-wise attention of self-features and spatial-wise attention of cooperation features are utilized simultaneously to generate refined cross features, where the cooperation features are viewed as a guidance to recalibrate the image self-features.

The details of the CFCM are shown in the bottom-left of Fig. 1. Specifically, a residual dense block (RDB) [45] is first applied to extract the hierarchical features of reference and target images. Then, to preserve the image pairs themselves information, the channel-wise attention of image itself is element-wise multiplied with the hierarchical features to obtain the self-features of reference and target images, respectively. In addition, to capture the cooperation relationship between reference and target images, the concatenated features are element-wise multiplied with the spatial-wise attention of concatenated features to obtain the cooperation features. Finally, the cooperation features of two views are element-wise summed with the self-features of respective view to obtain the refined cross features of reference and target images, respectively. The channel-wise attention is obtained by a global average pooling layer followed by two convolution layers and a sigmoid function, and the spatial-wise attention is obtained by two convolution layers and a sigmoid function. The refined cross features F^R_{out} and F^T_{out} of reference and target images are expressed as follows.

$$\begin{aligned} F^R_{out} &= F^{CR} \oplus F^S \\ F^T_{out} &= F^{CT} \oplus F^S \end{aligned} \tag{1}$$

with

$$\begin{cases} F^{CR} = RDB(F^R_{in}) \otimes M^R \\ F^{CT} = RDB(F^T_{in}) \otimes M^T \\ F^S = F^O \otimes M^O \end{cases} \tag{2}$$

where F^{CR} and F^{CT} denote reference image self-feature and target image self-feature, F^S denotes the cooperation features, F^R_{in} and F^T_{in} denote input reference image feature and input target image feature. F^O is the concatenated features, M^R and M^T denote reference image channel-wise attention and target image channel-wise attention, M^O is the spatial-wise attention of concatenated features, $RDB(\cdot)$ represents the residual dense block, $\oplus$ denotes element-wise summation, and $\otimes$ denotes element-wise multiplication. After that, the refined cross features obtained by each CFCM are progressively fed into a CCL to generate the corresponding progressive deep homography, which is further applied to warp the input images into aligned images by a spatial transformer module.

Interwoven image composition

After aligning the input image pairs, the aligned image pairs should be fused into a naturalness panorama image. On the one hand, since the overlapping regions in aligned images should be the same, the features in the overlapping regions need to be fused to show the common salient content. On the other hand, the non-overlapping regions in aligned images are different from each other, the features in the non-overlapping regions need to be preserved to emphasize their respective content. Existing swin transformer networks [46] utilize the shifted windows multi-head self-attention to capture the global important features of image pairs themselves, but ignore the complementary information between reference and target images. Thus, an interwoven swin transformer fusion sub-network is proposed to fuse the aligned image pairs into a panorama image. Specially, instead of independently integrating the features of aligned image pairs which ignore the connections of aligned pairs [47], an interaction guidance way with an attention interaction between aligned images can facilitate the complementary integration of global and local features from overlapping regions and forcing the retention of the original content and structure from non-overlapping regions.

As shown in Fig. 1, an interwoven swin transformer fusion sub-network is devised to effectively integrate the common global features and respective local features, which consists of two convolutional layers, four successive ISTMs, a concatenation layer, and a reconstruction layer. Specifically, four successive ISTMs aim to assist the fusion network to pay attention to the global features and local complementary information. More importantly, in each ISTM, an attention interwoven block (AIB) is designed between classic swin transformer to further capture the local complementary information between reference and target images. Since the network architectures of reference and target branches are symmetric, the details of AIB in the target image branch will be introduced as an example in the followings.

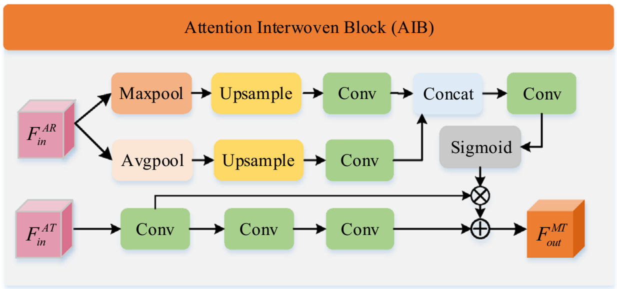

Fig. 2 The details of the attention interwoven block (AIB)

As shown in Fig. 2, for the aligned target image branch, the max-pooling and avg-pooling are first applied to the aligned reference features, which preserve the global important features and local detailed features simultaneously. Afterwards, the up-sampled features are fed into convolution layers separately and contacted together. Then, the concatenation features are sent into a convolution layer and a sigmoid function to generate the weighting map, which is further applied to provide complementary information to refine the aligned target features via the element-wise multiplication. Finally, the original aligned target features followed by three convolution layers are element-wise summation with the refined aligned target features. Specially, the formulas are defined as follows.

$$\begin{aligned}
H^R_m &= conv(up(max(F^{AR}_{in}))) \\
H^R_a &= conv(up(avg(F^{AR}_{in}))) \\
W^R &= \sigma(conv(concat(H^R_m, H^R_a))) \\
H^T &= conv(conv(conv(F^{AT}_{in}))) \\
F^{MT}_{out} &= H^T \oplus W^R \otimes conv(F^{AT}_{in})
\end{aligned} \tag{3}$$

where F^{MT}_{out} is the output interwoven features of AIB, F^{AR}_{in} and F^{AT}_{in} are the input aligned reference image feature and input aligned target image feature, H^R_m and H^R_a are the max-pooling features and avg-pooling features of aligned reference image, W^R is the weighting map, H^T is the enhanced feature map of aligned target image, σ is sigmoid function.

Following the ISTMs, in order to preserve original content and structure of non-overlapping regions between aligned images, the interwoven features aggregated by each ISTM of reference and target branches are first added to generate the final features, respectively. Then, the final features of reference and target images are concatenated in the channel dimension to obtain the fused deep features. Finally, a reconstruction layer [48] is applied to generate the final stitched image I^F, which can be formulated as:

$$I^F = Res(conv(concat(F^R_A, F^T_A))) \tag{4}$$

with

$$\begin{cases} F_A^R = \sum_{k-1}^{K} F_{Ik}^R \\ F_A^T = \sum_{k-1}^{K} F_{Ik}^T \end{cases} \tag{5}$$

where F_A^R and F_A^T are the final features of reference and target images, F_{Ik}^R and F_{Ik}^T denote the interwoven features of reference and target image, $Res(\cdot)$ is the reconstruction layer, and K is the number of ISTM.

Loss function

In this work, the proposed PAIC-Net aims to obtain a naturalness panorama by aligning and fusing the input image pairs. Therefore, an alignment loss and a composition loss are introduced to reduce alignment distortions and fusion artifacts.

Alignment loss To realize accurate alignment between reference and target images, an alignment loss is employed to improve the performance of image stitching. Specifically, we leverage the $L2$-norm to constrain the overlapping regions between reference and target images to be consistent at pixel level. The alignment loss L_A is defined as:

$$\begin{aligned} L_A = \| I^R \times \Im(I^X, H) - \Im(I^T, H)\|_2 \\ + \| I^T \times \Im(I^Y, H^{-1}) - \Im(I^R, H^{-1})\|_2 \end{aligned} \tag{6}$$

where I^R and I^T are the input reference and target images, $\Im(\cdot, \cdot)$ is the warping operation that is realized by spatial transformer network, I^X and I^Y are the all-one matrix with the same resolution as I^R and I^T, H denotes the estimated progressive homography matrix, and $\| \cdot \|_2$ denotes the $L2$-norm.

Composition loss Most image stitching methods only consider the pixel consistency during image alignment but ignore the structure and texture consistency during the image composition. In order to align image pairs accurately and prevent the seam distortions of the overlapping regions simultaneously, a composition loss is designed to improve the seam naturalness, which includes a structure loss and a texture loss. More specifically, to force the stitched image to preserve the structure information, a structure loss is utilized to the measure the structural similarity between stitched image and input image pairs. Meanwhile, a seamless stitched image should have similar texture details to the input image pairs, so a texture loss is utilized to reduce texture details changes. The composition loss L_C is defined as:

$$L_C = L_{stru} + L_{text} \tag{7}$$

with

$$\begin{cases} L_{stru} = (1 - ssim(O^F, O^{AR})) + (1 - ssim(O^F, O^{AT})) \\ L_{text} = \frac{1}{MN} \| |\nabla O^F| - max(|\nabla O^{AR}|, |\nabla O^{AT}|) \|_1 \end{cases} \tag{8}$$

where L_{stru} and L_{text} are the structure loss and the texture loss, O^{AR} and O^{AT} are image patches extracted at a pixel local O from aligned reference and target images, O^F is the image patch extracted from the final stitched image at same location O. M and N represent the horizontal and vertical directions, $ssim(\cdot, \cdot)$ denotes structural similarity operation, $|\cdot|$ and $\|\cdot\|_1$ are absolute operator and the $L1$-norm, $\bigtriangledown$ denotes the *Sobel* gradient operation, and $max(\cdot)$ is the element-wise maximum selection.

Finally, we sum the alignment loss L_A and composition loss L_C to obtain the final object function, which is derived as follows.

$$L_{total} = \lambda_1 L_A + \lambda_2 L_C \tag{9}$$

where λ_1 and λ_2 are factors for the alignment loss and composition loss.

Experiments

In this section, the experimental setup of proposed framework is introduced firstly. Then, we present the comparisons with SOTA methods to demonstrate the superior performance of proposed PAIC-Net. The ablation studies are conducted finally to estimate the contribution of different components.

Experimental setup

Datasets We evaluate the performance of the proposed PAIC-Net on two datasets. Two public image stitching datasets, i.e. UDIS-D dataset [11] and PTIS dataset [5], are utilized to conduct the quantitative and qualitative experiments. The UDIS-D dataset contains 10,440 image pairs for training and 1106 image pairs for testing with different overlapping rates. It includes variable scenarios, mainly originating from moving videos. The PTIS dataset contains 35 challenging image pairs with large parallax for testing. Most of PTIS dataset were captured from real-world outdoor scenarios in different resolution. In addition, some classic image stitching instances from [8] are also compared to illustrate the robust of the proposed PAIC-Net.

Implementation details To obtain a panorama image with good visual effect, we train the proposed PAIC-Net in two stages. In the first stage, we train the progressive homography alignment module by the $ADAM$ optimizer method

[49] for up to 100 epochs, where an initial value of 0.0002 is applied to train the deep homography module. In the second stage, we train the deep image stitching module with the parameters of the progressive homography alignment module being fixed. The training strategy is the same as that of the progressive homography alignment module. The batch size and momentum are set to 4 and 0.9, respectively. In addition, some data preprocessing methods is applied on the deep image stitching network, e.g., image denoising is used to reduce the impact of noise on the image stitching results. Meanwhile, some data augmentation methods are utilized to enhance the illumination robustness of deep image stitching network, such as adding random brightness transformation into the training dataset. In our experiments, the size of image patch O^{AR}, O^{AT} and O^{F} in the composition loss is set as 13×13, the size of sliding window of SSIM is set as 8×8, and the step size is set as 1. Specifically, when calculating SSIM, image patches are first divided into multiple blocks by the sliding window to compare the SSIM of each two blocks. After traversing the entire image patch, the average values of all blocks are taken as the SSIM of image patches O^{AR} and O^{F}, and the SSIM of image patches O^{AR} and O^{F}. The λ_1 and λ_2 are set to 0.3 and 0.3 after parameter adjustment. After many simulations and experiments, these were the optimum parameters. The implementation of the deep image stitching model is based on Pytorch and the deep model training is conducted on a single GPU with NVIDIA GeForce RTX 2080Ti.

Evaluation metrics For quantitative evaluations of the deep image stitching, two commonly-used metrics, i.e. SSIM metric [50] and PSNR metric [51], are adopted in this section. Specifically, the SSIM metric of the overlapping regions is applied to measure the similarity between two image pairs. The higher the SSIM metric is, the better the visual quality achieves. Meanwhile, the PSNR metric of the overlapping regions is also applied to evaluate the degree of image distortions during the stitching process. A higher PSNR metric means a higher-quality panorama image with less distortions. These evaluation metrics are complementary and can provide a comprehensive evaluation.

Compared with SOTA methods

In this section, we compare the proposed deep image stitching network with other eight other SOTA methods, i.e., APAP [8], adaptive as-natural-as-possible warp (AANAP) [23], natural image stitching (NIS) [24], robust elastic warping (REW) [9], deep homography estimation (DHE) [32], unsupervised deep image stitching (UDIS) [11], and edge-preserved image stitching (EPIS) [20], image stitching with manifold optimization (ISMO) [25]. APAP [8], AANAP [23], NIS [24], REW [9] and ISMO [25] are traditional methods, while DHE [32], UDIS [11] and EPIS [20] are deep learning methods. To obtain the results for fair comparison, we use the code provided by the authors under the default parameters or the figures published in the papers.

Qualitative comparison evaluation

To evaluate the performance of the proposed network, several qualitative comparison results obtained by different image stitching methods for visual comparison are presented in Fig. 3. The top four images are selected from the UDIS-D dataset while the middle four images are selected from the PTIS dataset, and the last two image pairs from the classic image stitching dataset. These images are representative and challenging test cases because the scenes contain a large number objects and structures, which makes them complex. Some of these scenes are captured at different shooting positions with large parallax between reference and target images, which can be utilized to evaluate the performance of image alignment. In addition, some test images with differences in brightness and color are chosen to estimate the performance of image composition. For space limitations, only the image stitching results of some representative scenes and the challenges scenes are presented in this section. In addition, similar to the training data, we also perform image denoising preprocessing on the test image to eliminate the impact of image noise on the image stitching results.

From Fig. 3, we can see that some classic image stitching methods, i.e. APAP [8] and AANAP [23] generate some misalignment in the overlapping regions in most instances, because these two methods only utilize local parametric warps and ignore the global homography estimation. For example, the pillar in the third row has obvious misalignment and the steel frame in the ninth row has ghosting in Fig. 3b, c. In addition, NIS [24] and REW [9] obtain more natural images by both considering the global and local deformation warps, but they are incapable of aligning images with severe occlusions. As one of the latest traditional image stitching methods, ISMO [25] obtain a visually good image with less shape distortions by introducing spatially varying manifold homographies. However, it does not consider the appearance difference which may be inconsistent in the overall intensity.

Compared with traditional methods, these two deep learning methods, i.e. DHE [32] and EPIS [20], show better panorama with a few slight artifacts, but some edges discontinuities are appeared in the overlapping joints. For instance, the leaves in the fourth row have some distorted edges in Fig. 3f, h. Furthermore, the UDIS [11] shows better stitched images with fewer parallax artifacts, but when the parallax increases, the alignment performance would be decreased. On the contrary, we can observe that the proposed PAIC-Net provides naturalness images stitching results without ghosting effects, especially for some scenes

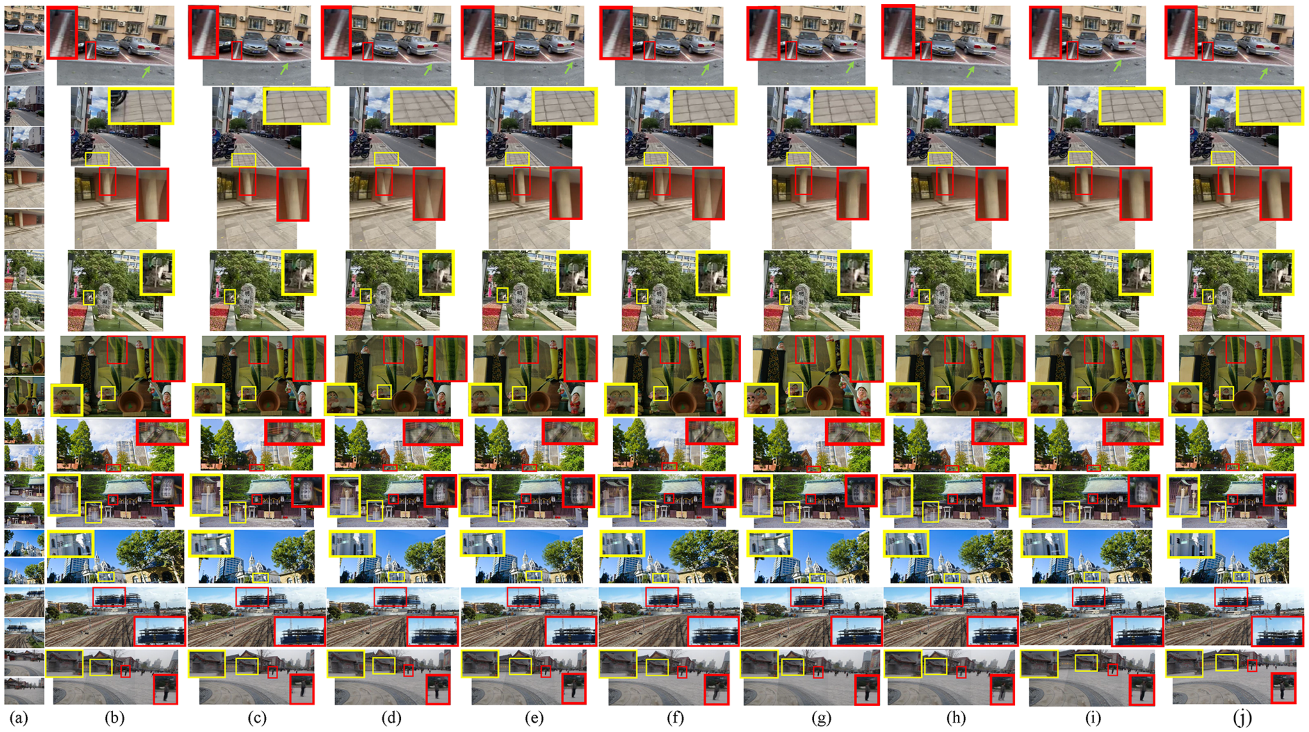

Fig. 3 Visual comparison of the image stitching quality between different image stitching methods. From left to right: **a** the input images, **b** APAP [8], **c** AANAP [23], **d** NIS [24], **e** REW [9], **f** DHE [32], **g** UDIS [11], **h** EPIS [20], **i** ISMO [25], **j** the proposed method

with large parallax. This actually verifies the effectiveness of the proposed progressive homography alignment which progressively integrate self-features and cooperation features and the proposed interwoven swin transformer fusion which effectively fuses the local complementary information. Benefiting from the combination of these two stages, the proposed method obtains promising performance of stitching images.

Quantitative comparison evaluation

To further demonstrate the performance of proposed PAIC-Net comprehensively, we also conduct quantitative evaluation of different methods. Table 1 illustrates comparison results in terms of SSIM and PSNR. As shown in Table 1, it shows that the PAIC-Net outperforms the other image stitching methods in the vast majority of cases. More specifically, all traditional image stitching methods obtain lower SSIM and PSNR results than the proposed method. It is because these methods rely on the feature points extraction and matching, where the error extraction and mismatch of feature points affect homography estimation accuracy. For instances, for the UDIS-D dataset, SSIM of PAIC-Net is 0.3175 higher than APAP [8], and its PSNR is 5.3504 higher than APAP [8]. Similarly, for the PTIS dataset, SSIM of PAIC-Net is 0.0132 higher than ISMO [25], and its PSNR is 0.9301 higher than ISMO [25].

Table 1 Quantitative comparison on the UDIS-D dataset and PTIS dataset

Methods	UDIS-D		PTIS	
	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑
APAP	0.5545	18.5968	0.4533	16.5894
AANAP	0.5032	20.4958	0.6752	18.4946
NIS	0.6883	18.0592	0.6101	21.1298
REW	0.7854	20.4938	0.765	22.489
DHE	0.7658	18.5839	0.6492	20.2277
UDIS	0.8653	23.6849	0.7659	21.7269
EPIS	0.7896	22.4872	0.711	20.5202
ISMO	0.7996	23.5081	0.7734	21.2809
PAIC-Net	**0.872**	**23.9472**	**0.7866**	**22.211**

Bold indicates the best performance

By comparison, two supervised image stitching methods (i.e., DHE [32] and EPIS [20]) are obviously improved on most test images in terms of SSIM and PSNR. However, these two methods cannot achieve favorable and stable performance due to the simple mapping relationship estimation. For example, for the UDIS-D dataset, SSIM of PAIC-Net is promoted by 0.1062 comparing with DHE [32] and 0.0824 comparing with EPIS [20]. Furthermore, UDIS [11] deliver a slightly inferior performance to the proposed network. It suffers from performance degradation, because the scenes with large parallax to be reconstructed. For instances, for

Table 2 Ablation studies on the influence of the CFCM in the progressive homography alignment

Cases	UDIS-D		PTIS	
	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑
w CFCM-V0	0.7085	17.6958	0.4321	15.889
w CFCM-V1	0.7849	18.795	0.6452	17.794
w CFCM-V3	**0.872**	**23.9472**	**0.7866**	**22.211**
w CFCM-V5	0.7232	21.596	0.5643	20.657

Bold indicates the best performance

the PTIS dataset, SSIM of PAIC-Net is 0.0207 higher than UDIS [11], and its PSNR is 0.4841 higher than UDIS [11]. Benefiting from the proposed alignment and fusion modules, the PAIC-Net can well handle these challenging images in reducing artifacts and discontinuity.

Ablation studies

In this section, we perform the ablation studies on the network components and loss function. Specifically, the effectiveness of deep alignment stage and the deep composition stage are tested firstly. Then, the alignment loss and composition loss are validated in details.

Ablation study on the CFCM in the progressive homography alignment

To verify the contribution of CFCM in the progressive homography alignment network, the ablation studies are conducted in this section. Specifically, we increase the numbers of CFCMs and denote the deep network with M CFCMs as CFCM-VM, where $M \in \{0, 1, 3, 5\}$. The evaluation experiment results are shown in Table 2. From the Table 2, we can see that the PAIC-Net outperforms the case without all CFCMs (w CFCM-V0) in the quantitative comparison, i.e. for the UDIS-D dataset, the SSIM of the proposed PAIC-Net is 0.1635 higher than the case without CFCM, and its PSNR is 6.2514 higher than the case without CFCM. This implies the necessity of CFCM in the progressive homography alignment. Meanwhile, it can be seen that the stitching performance of PAIC-Net is improved with the increase of CFCM, but when CFCM exceeds 3, the SSIM and PSNR of PAIC-Net decrease. Therefore, we set $M = 3$ to achieve a good model performance.

To facilitate comparison, Fig. 4 presents several panorama images produced by the case without all CFCMs (w CFCM-V0) and the proposed PAIC-Net (w CFCM-V3), respectively. As shown in the Fig. 4c, the proposed image stitching network can better align the foreground objects and the background reliably due to the accurate feature extraction ability of CFCM.

Ablation study on the ISTM in the interwoven image composition

The ablation studies of ISTM in the interwoven image composition are performed. To be specific, the numbers of ISTMs are gradually increased and denote the deep model with K ISTMs as ISTM-VK, where $K \in \{0, 2, 4, 6\}$. The evaluation experiment results are shown in Table 3, we can observe that the deep PAIC-Net achieves the worst results when all ISTMs are removed (w ISTM-V0). In addition, it can be seen that the performance of PAIC-Net is improved with the increase of ISTM, but we also notice that when ISTM exceeds 4, the performance of PAIC-Net decreases. Therefore, we set $K = 4$ to obtain a good SSIM and PSNR metrics.

Moreover, we also conduct the comparative experiments between AIB in the ISTM and commonly used spatial-wise attention (w SA) [52] / channel-wise attention modules [53] (w CA) in the ablation study. We use the classic swin transformer [46] as a substitute for the interwoven swin transformer (w/o AIB). From Table 4, we can observe that the case without AIB gives the worst SSIM and PSNR metrics. This indicates that the necessity of using AIB for image composition. Meanwhile, the case with spatial-wise attention and the case with channel-wise attention can not effectively improving image stitching performance in terms of SSIM and PSNR metrics.

Furthermore, we also provide some visual comparisons in Fig. 5 to intuitively exhibit the advantages of the proposed method in deep image composition. As one can see, our fusion module (ISTM-V4) creates more natural-looking stitching results. For instance, there is obvious color inconsistency and misalignment in the red and yellow boxes in Fig. 5b, but the PAIC-Net provides high-quality stitched image results without seam artifacts.

Ablation study on loss function

To understand the effectiveness of each loss in the total loss function, we also conduct the ablation studies in this section. The total loss function includes the alignment loss and the composition loss. In the experiments, we remove one loss function and evaluate the performances by keeping the other loss function unchanged. The SSIM and PSNR metrics and qualitative comparisons are represented in Table 5 and in Fig. 6. From Table 5, compared with the complete total loss function, any of these loss functions removed would lead to the performance degradation problem. In addition, from the visual comparisons in Fig. 6, we observe that the proposed method can reduce deformation distortions and avoid seam artifacts in the panorama image in different image dataset.

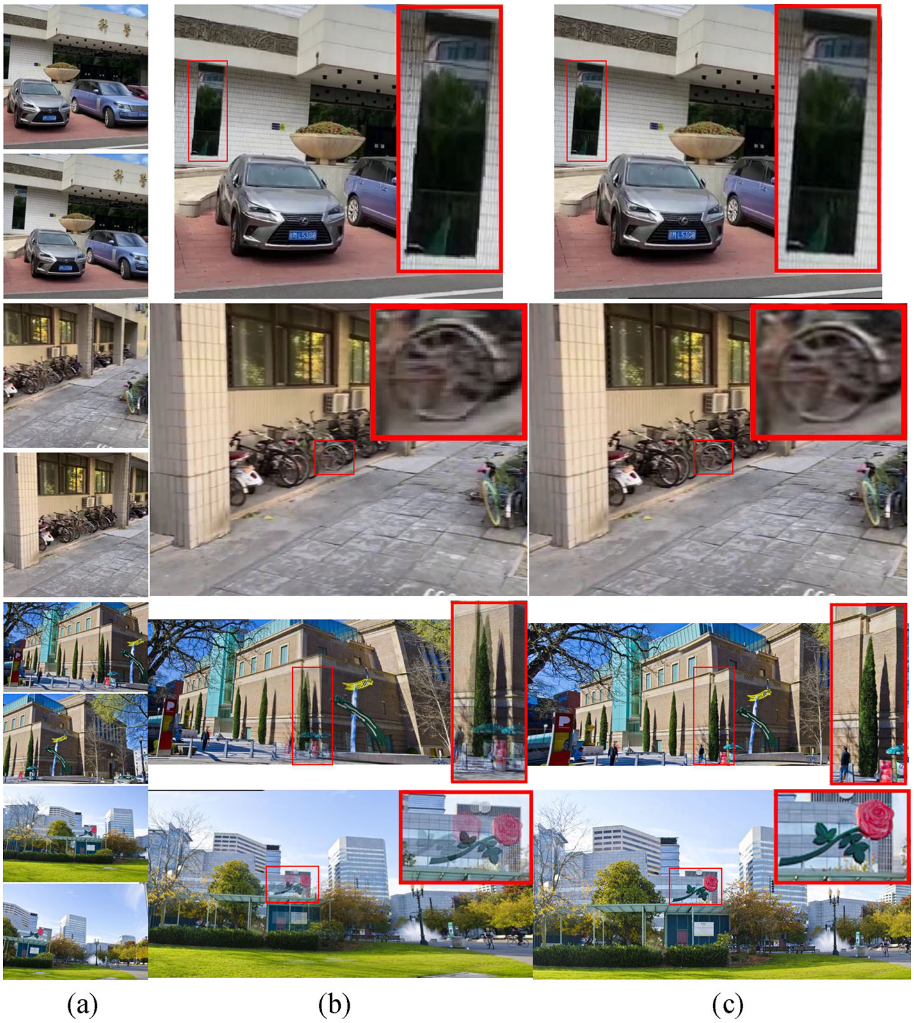

Fig. 4 Qualitative comparison results of the PAIC-Net with/without CFCM. From first to second: instances from the UDIS-D dataset. From third to fourth: instances from PTIS dataset. **a** The input images, **b** results without CFCM, **c** the proposed method

Table 3 Ablation studies on the influence of the ISTM in the interwoven image composition

Cases	UDIS-D		PTIS	
	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑
w ISTM-V0	0.6743	20.471	0.6578	18.794
w ISTM-V2	0.7785	22.642	0.7256	20.684
w ISTM-V4	**0.872**	**23.9472**	**0.7866**	**22.211**
w ISTM-V6	0.7835	21.064	0.7559	20.456

Bold indicates the best performance

Table 4 Ablation studies on the influence of the AIB in the interwoven swin transformer fusion

Cases	UDIS-D		PTIS	
	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑
w/o AIB	0.7257	19.9842	0.6391	17.009
w SA	0.7491	21.345	0.7286	18.074
w CA	0.7932	21.653	0.7053	20.546
PAIC-Net	**0.872**	**23.9472**	**0.7866**	**22.211**

Bold indicates the best performance

Ablation study on different loss weights

Finally, we conduct the ablation experiments to evaluate the stitching performance with different loss factors λ_1 and λ_2 in the Eq. (9). Figure 7 shows the SSIM metric of final loss function with different factors. To be specific, λ_1 is first set as 0 to obtain the best λ_2. From the Fig. 7a, the optimal SSIM is obtained when λ_2 is set to 0.3. Finally, λ_2 is fixed as 0.3 to search for the best λ_1. As illustrated in Fig. 7b, the best SSIM metric is obtained when λ_1 set to 0.3.

Computational complexity and discussions

To analyze the complexity of the PAIC-Net, the computational complexity of different image stitching methods is compared and discussed. The processing environment has a single GPU with NVIDIA GeForce RTX 2080Ti. Table 6 shows the computational complexity comparisons between different image stitching methods on UDIS-D dataset and PTIS dataset. Specifically, the running time of nine image

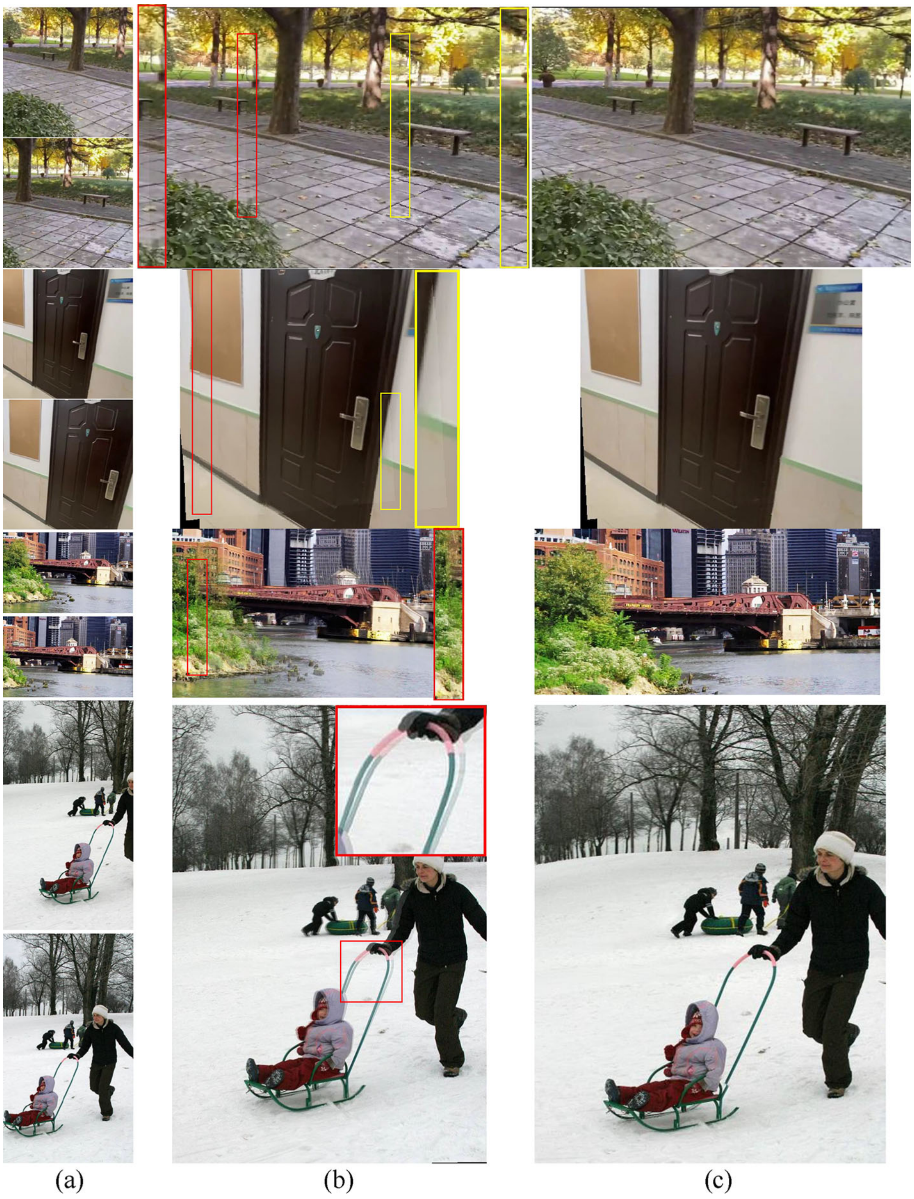

Fig. 5 Qualitative comparison results of the proposed method with/without ISTM. From first to second: instances from the UDIS-D dataset. From third to fourth: instances from PTIS dataset. **a** The input images, **b** results without ISTM, **c** the proposed method

Table 5 Ablation study on the loss function

Cases	UDIS-D		PTIS	
	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑
w/o alignment loss	0.5914	17.5829	0.5914	17.5829
w/o composition loss	0.7381	19.583	0.7381	19.583
PAIC-Net	**0.872**	**23.9472**	**0.7866**	**22.211**

Bold indicates the best performance

stitching methods are calculated, and the number of floating-point operations (FLOPs) of four deep learning-based image stitching methods is also compared. Compared with the traditional image stitching methods, the deep learning-based methods have a significant improvement due to the GPU acceleration strategy. However, compared with other deep learning methods, the proposed method costs a little more running time and FLOPs, because it requires to use the interwoven swin transformer based on self-attention to learn the correspondence relationship between reference image and target image. Despite this, the proposed method provides higher-quality image stitching results both quantitatively and qualitatively.

Limitations and future work

The proposed PAIC-Net obtains superior performance by combining a progressive homography alignment module and an interwoven swin transformer fusion module. However,

Fig. 6 Qualitative comparison results of the proposed method with/without different loss. The first instance from the UDIS-D dataset, the second instance from PTIS dataset, the third instance from [8]. **a** The input images, **b** results without alignment loss, **c** results without composition loss, **d** the proposed method

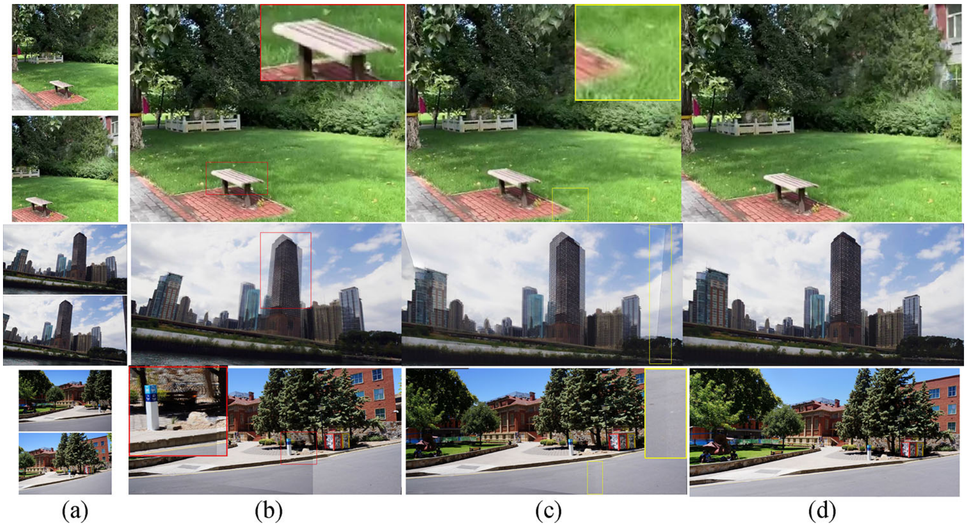

Table 6 Computational complexity between different image stitching methods on the UDIS-D dataset and PTIS dataset

Methods	UDIS-D		PTIS		Average time (s)	Platform
	Time (s)	FLOPs (G)	Time (s)	FLOPs (G)		
APAP	2.907	–	4.861	–	3.884	Matlab
AANAP	6.368	–	5.494	–	5.931	Matlab
NIS	12.863	–	18.321	–	15.592	Matlab
REW	9.63	–	12.348	–	10.989	Matlab
DHE	0.327	51.674	0.109	29.786	0.218	Pytorch
UDIS	0.657	152.897	0.233	127.435	0.445	TensorFlow
EPIS	0.456	135.456	0.162	103.652	0.309	TensorFlow
ISMO	10.796	–	6.854	–	8.825	Matlab
PAIC-Net	0.687	175.678	0.469	155.531	0.578	Pytorch

we also observe some limitations of the proposed method in practical applications. Firstly, the proposed method utilizes the interwoven swin transformer based on self-attention to learn the correspondence relationship between reference image and target image, which may introduce more running time to generate real panorama images. Secondly, the proposed method assumes that the transformation between reference image and target image is a global deep homography, which may not suitable for casually captured image pairs in actual scenes. For future work, we plan to investigate a lightweight image stitching network based on knowledge distillation that can achieve a better balance between running time and model size, and model performance. In addition, we would like to estimate the local deep homography to transform reference image and target image.

Conclusion

This paper proposes a novel progressive alignment and interwoven fusion network for image stitching. Firstly, a progressive homography alignment module is presented to align images, which progressively warp the reference and target images by leveraging the channel-wise attention of self-features and spatial-wise attention of cooperation features. Secondly, an interwoven swin transformer fusion module is exploited to effectively fuse aligned image pairs, where the complementary information is learned in an interweaved way. Finally, an alignment loss and a composition loss are applied to eliminate image alignment distortions and enhance seam consistency. Experiment results demonstrate the superiority of the proposed image stitching method outperforms SOTA solutions.

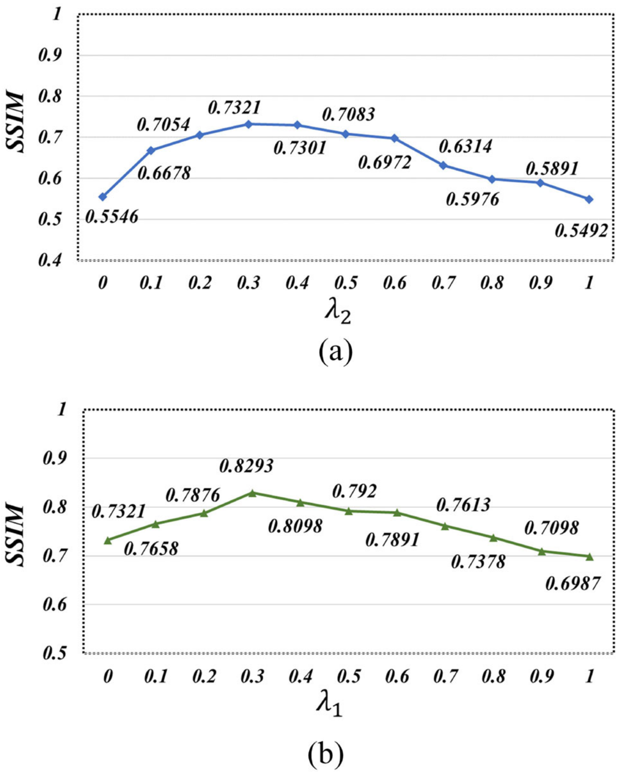

Fig. 7 SSIM metric for the proposed method with different loss factors

Acknowledgements This work was supported in part by the Natural Science Foundation of Tianjin under Grant No. 22JCQNJC00010, in part by the Scientific Research Project of Tianjin Educational Committee under Grant No. 2022KJ011, and in part by the National Natural Science Foundation of China under Grant No. 62171321 and No. 62471335.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Madhusudana PC, Soundararajan R (2019) Subjective and objective quality assessment of stitched images for virtual reality. IEEE Trans Image Process 28(11):5620–5635
2. Luo X, Li Y, Yan J, Guan X (2020) Image stitching with positional relationship constraints of feature points and lines. Pattern Recogn Lett 135:431–440
3. Delphin DA, Bhatt MR, Thiripurasundari D (2021) Holoentropy measures for image stitching of scenes acquired under CAMERA unknown or arbitrary positions. J King Saud Univ Comput Inf Sci 33(9):1096–1107
4. Tian C, Shao F, Chai X, Jiang Q, Xu L, Ho Y-S (2023) Viewport-sphere-branch network for blind quality assessment of stitched 360° omnidirectional images. IEEE Trans Circuits Syst Video Technol 33(6):2546–2560
5. Zhang F, Liu F (2014) Parallax-tolerant image stitching. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Columbus, OH, USA, pp 3262–3269
6. Brown M, Lowe D G (2003) Recognising panoramas. In: Proceedings of the IEEE international conference on computer vision, Nice, France, pp 1218–1225
7. Gao J, Kim S J, Brown M S (2011) Constructing image panoramas using dual-homography warping. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Colorado Springs, CO, USA, pp 49–56
8. Zaragoza J, Chin T-J, Tran Q-H, Brown MS, Suter D (2014) As-projective-as-possible image stitching with moving DLT. IEEE Trans Pattern Anal Mach Intell 36(7):1285–1298
9. Li J, Wang Z, Lai S, Zhai Y, Zhang M (2018) Parallax-tolerant image stitching based on robust elastic warping. IEEE Trans Multimed 20(7):1672–1687
10. Li A, Guo J, Guo Y (2021) Image stitching based on semantic planar region consensus. IEEE Trans Image Process 30:5545–5558
11. Nie L, Lin C, Liao K, Liu S, Zhao Y (2021) Unsupervised deep image stitching: reconstructing stitched features to images. IEEE Trans Image Process 30:6184–6197
12. Nie L, Lin C, Liao K, Liu S, Liu M, Zhao Y (2020) A view-free image stitching network based on global homography. J Vis Commun Image Represent 73(102950):1–9
13. Nie L, Lin C, Liao K, Liu S, Zhao Y (2022) Deep rectangling for image stitching: a learning baseline. In: Proceedings of the IEEE conference on computer vision and pattern recognition, New Orleans, LA, USA, pp 5730–5738
14. Kweon H, Kim H, Kang Y et al (2023) Pixel-wise warping for deep image stitching. In: The thirty-seventh AAAI conference on artificial intelligence, vol 37, no 1, pp 1196–1204
15. Song D-Y, Um G-M, Lee HK, Cho D (2021) End-to-end image stitching network via multi-homography estimation. IEEE Signal Process Lett 28:763–767
16. Shao R, Wu G, Zhou Y, Fu Y, Fang L, Liu Y (2021) LocalTrans: a multiscale local transformer network for cross-resolution homography estimation. In: Proceedings of the IEEE international conference on computer vision, Montreal, QC, Canada, pp 14870–14879
17. Liu S, Lu Y, Jiang H, Ye N, Wang C, Zeng B (2023) Unsupervised global and local homography estimation with motion basis learning. IEEE Trans Pattern Anal Mach Intell 45(6):7885–7899
18. Nie L, Lin C, Liao K, Liu S, Zhao Y (2022) Depth-aware multi-grid deep homography estimation with contextual correlation. IEEE Trans Circuits Syst Video Technol 32(7):4460–4472
19. Cao S -Y, Hu J, Sheng Z, Shen H -L (2022) Iterative deep homography estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, New Orleans, LA, USA, pp 1869–1878

20. Nie L, Lin C, Liao K et al (2022) Learning edge-preserved image stitching from multi-scale deep homography. Neurocomputing 491:533–543
21. Dai Q, Fang F, Li J, Zhang G, Zhou A (2021) Edge-guided composition network for image stitching. Pattern Recogn 118(108019):1–13
22. Lin W -Y, Liu S, Matsushita Y, Ng T -T, Cheong L -F (2011) Smoothly varying affine stitching. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Colorado Springs, CO, USA, pp 345–352
23. Lin C-C, Pankanti SU, Ramamurthy KN, Aravkin AY (2015) Adaptive as-natural-as-possible image stitching. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, MA, USA, pp 1155–1163
24. Chen Y-S, Chuang Y-Y (2016) Natural image stitching with the global similarity prior. In: Proceedings of the European conference on computer vision, Amsterdam, Netherlands, pp 186–201
25. Zhang L, Huang H (2023) Image stitching with manifold optimization. IEEE Trans Multimed 25:3469–3482
26. Liao T, Li N (2020) Single-perspective warps in natural image stitching. IEEE Trans Image Process 29:724–735
27. Zhang Z, Yang X, Xu C (2023) Natural image stitching with layered warping constraint. IEEE Trans Multimed 25:329–338
28. Xue W, Xie W, Zhang Y, Chen S (2022) Stable linear structures and seam measurements for parallax image stitching. IEEE Trans Circuits Syst Video Technol 32(1):253–261
29. Zhang Y, Lai Y-K, Zhang F-L (2021) Content-preserving image stitching with piecewise rectangular boundary constraints. IEEE Trans Vis Comput Graph 27(7):3198–3212
30. Yan N, Mei Y, Xu L et al (2023) Deep learning on image stitching with multi-viewpoint images: a survey. Neural Process Lett 55(4):3863–3898
31. Zhang J, Wang C, Liu S, Jia L, Wang J, Zhou J, Sun J (2020) Content-aware unsupervised deep homography estimation. In: Proceedings of the European conference on computer vision, pp 1–16
32. Zhao Q, Ma Y, Zhu C et al (2021) Image stitching via deep homography estimation. Neurocomputing 450:219–229
33. Nie L, Lin C, Liao K, Liu S, Zhao Y (2023) Parallax-tolerant unsupervised deep image stitching. In: Proceedings of the international conference on computer vision, Paris, France, pp 7399–7408
34. Xie M, Sun B (2023) Fast image reconstruction network in image stitching. Optoelectron Lett 19(20):635–640
35. Khamiyev I, Issa D, Akhtar Z, Demirci MF (2023) Panoramic image generation using deep neural networks. Soft Comput 27:8679–8695
36. Du K, Wang Y, Yin J, Cao L, Guo Y (2024) RelTransformer: a transformer-based long-tail visual relationship recognition. Complex Intell Syst. https://doi.org/10.1007/s40747-024-01456-6
37. Wang D, Zhang J, Du B, Zhang L, Tao D (2023) DCN-T: dual context network with transformer for hyperspectral image classification. IEEE Trans Image Process 32:2536–2551
38. Yang R, Liu K, Xu S, Yin J, Zhang Z (2024) ViT-UperNet: a hybrid vision transformer with unified-perceptual-parsing network for medical image segmentation. Complex Intell Syst 10:3819–3831
39. Zheng S et al (2021) Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Nashville, TN, USA, pp 6877–6886
40. Gao G, Xu Z, Li J, Yang J, Zeng T, Qi G-J (2023) CTCNet: a CNN-transformer cooperation network for face image super-resolution. IEEE Trans Image Process 32:1978–1991
41. Li Y, Yao T, Pan Y, Mei T (2023) Contextual transformer networks for visual recognition. IEEE Trans Pattern Anal Mach Intell 45(2):1489–1500
42. Esser P, Rombach R, Ommer, B (2021) Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Nashville, TN, USA, pp 12868–12878
43. Li J, Zhu J, Li C, Chen X, Yang B (2022) CGTF: convolution-guided transformer for infrared and visible image fusion. IEEE Trans Instrum Meas 71(5012314):1–14
44. Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K (2015) Spatial transformer networks. Neural information processing systems, pp 2017–2025
45. Zhang Y, Tian Y, Kong Y, Zhong B, Fu Y (2021) Residual dense network for image restoration. IEEE Trans Pattern Anal Mach Intell 43(7):2480–2495
46. Liu Z et al (2021) Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE international conference on computer vision, Montreal, QC, Canada, pp 9992–10002
47. Xie X et al (2023) MRSCFusion: joint residual swin transformer and multiscale CNN for unsupervised multimodal medical image fusion. IEEE Trans Instrum Meas 72(5026917):1–17
48. Ma J, Tang L, Fan F, Huang J, Mei X, Ma Y (2022) SwinFusion: cross-domain long-range learning for general image fusion via swin transformer. IEEE/CAA J Autom Sin 9(7):1200–1217
49. Kingma D P, Ba J (2014) Adam: a method for stochastic optimization. https://doi.org/10.48550/arXiv.1412.6980
50. Wang Z, Bovik A, Sheikh H, Simoncelli E (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13:600–612
51. Sheikh H, Sabir M, Bovik A (2006) A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Trans Image Process 15:3440–3451
52. Woo S, Park J, Lee J -Y, Kweon IS (2018) CBAM: convolutional block attention module. In: Proceedings of the European conference on computer vision, Munich, Germany, pp 1–17
53. Hu J, Shen L, Albanie S, Sun G, Wu E (2020) Squeeze-and-excitation networks. IEEE Trans Pattern Anal Mach Intell 42(8):2011–2023

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.