Neurocomputing 491 (2022) 533–543

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Learning edge-preserved image stitching from multi-scale deep homography

Lang Nie [a,b], Chunyu Lin [a,b,*], Kang Liao [a,b], Yao Zhao [a,b]

[a] With Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China
[b] Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 10044, China

ARTICLE INFO

Article history:
Received 8 June 2021
Revised 4 August 2021
Accepted 12 December 2021
Available online 17 December 2021

Keywords:
Image stitching
Homography estimation
Deep learning
Computer vision

ABSTRACT

Image stitching is a classical and challenging technique in computer vision, which aims to generate an image with a wide field of view. The traditional methods heavily depend on feature detection and require the feature points to be dense and evenly distributed in the image, leading to poor robustness in low-texture scenes. Learning methods are rarely studied due to the unavailability of ground truth stitched results, showing unreliable performance on real-world datasets. In this paper, we propose an image stitching learning framework, which consists of a multi-scale deep homography module and an edge-preserved deformation module. First, we design a multi-scale deep homography module to estimate the accurate homography progressively from coarse to fine. After that, an edge-preserved deformation module is designed to learn the deformation rules of image stitching from edge to content, generating the stitched image with artifacts eliminated. Besides, the proposed supervised learning framework can stitch images of arbitrary resolutions and demonstrate good generalization capability in real-world images. Experiments show that our superiority to the existing homography solutions and image stitching algorithms.

1. Introduction

Due to the limited field-of-view (FOV), a single photo may not be able to display the complete region of interest (ROI). To tackle this problem, a stitched image of a wider FOV can be obtained by stitching images from different viewing positions, which plays an important role in various applications such as autonomous driving [44,21], immersive communication [16], virtual reality (VR) [1,17].

Traditional image stitching methods follow similar steps: feature detection and matching, image registration, and image composition. These methods eliminate the ghosting effects caused by parallax by proposing a spatial adaptive warping model to align the contents [34,11,30,46,5,4,6,27,25,23,32,24,26] or searching an optimal seam to composite the stitched image [9,47,28,12]. However, the performance of these methods greatly depends on the number and distribution of hand-craft feature points, often leading to failures in low-texture scenarios.

The existing deep learning solutions are still in development. They achieve this technology by training an image stitching network on a synthetic stitching dataset in a supervised manner [37,50]. Deep learning solutions can work robustly in low-texture scenarios due to the robust deep features extracted by the neural network. However, the performance of this model on real-world datasets is unsatisfactory, and it cannot handle input of arbitrary resolution.

Considering those above traditional and learning methods' limitations, we propose a novel deep image stitching framework to stitch images of arbitrary resolutions from arbitrary shooting positions in a flexible learning wary. The proposed framework is composed of a multi-scale deep homography module and an edge-preserved deformation module. The first module achieves the homography estimation and image registration, and the remaining module stitches the images from edge to content.

In deep homography module, we found the following two common problems in the existing learning methods [8,36,48,22]: 1) Only the feature maps learned by the last convolutional layer are adopted to predict the homography, while they ignore the features of different scales learned by other convolutional layers. 2) Learning the matching relationship of features by convolutional layers is inefficient, making these methods fail to work in scenes of low overlap rates. In these methods, the receptive fields of convolu-

* Corresponding author at: With Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China.
E-mail address: cylin@bjtu.edu.cn (C. Lin).

https://doi.org/10.1016/j.neucom.2021.12.032
0925-2312/

tional layers are limited by the kernel size, while the distance between matched features can be much longer than it.

To address the above problems, we propose a feature-level multi-scale deep homography network. Specifically, we first adopt the feature pyramid to extract multi-scale features and every scale of features can contribute to the homography regression. Then we extract the feature matching through a scheme of feature correlation instead of convolutional layers. In particular, the feature matching is extracted from global to local, thus the homography can be estimated from coarse to fine.

Subsequently, we design an edge-preserved deformation module to stitch the warped images (Fig. 1 (b)) that can be obtained from the previous module. Traditional image fusion strategy eliminates the misalignments (shown in Fig. 1 (c)) by assigning adaptive weights to the warped images to hide the artifacts. Different from it, the proposed deformation module learns to composite the stitched image in two steps: 1) The network tends to remove the misalignments at the cost of edge discontinuity between the warped reference image and the non-overlapping areas of the warped target image (shown in Fig. 1 (d)). 2) Learn to correct discontinuity by pixel-level deformation from edge to content (shown in Fig. 1 (e)).

In experiments, we evaluate our method on the tasks of homography estimation and image stitching, demonstrating our robustness and efficacy. The contributions of this paper are summarized as follows:

- We design a multi-scale deep homography model, which integrates the feature pyramid and feature correlation simultaneously, enabling our robustness and efficacy in the scenes of low overlap rates.
- We propose an edge-preserved deformation network to stitch the warped images, eliminating the ghosting effects and keeping the edge continuity of the stitched image simultaneously.
- In the presence of fully connected layers, we designed a flexible mechanism combining image scaling and homography scaling to stitch images of arbitrary resolution.

2. Related work

2.1. Traditional image stitching

Traditional image stitching methods can be simply divided into the following two categories.

Spatial Adaptive Warping. Traditional schemes stitch images with a single global homography, causing noticeable ghosting effects [13]. To construct image panoramas with fewer artifacts, Gao *et al.* proposed a dual-homography method (DHW) to represent the warpings of the foreground and background, respectively [11]. To align different areas in the image domain, spatially adaptive warpings are calculated to stitch images as-projectively-as-possible (APAP) in the work of Zaragoza *et al.* [46]. Dividing pictures into dense grids, APAP calculates the spatially-adaptive warpings using moving DLT to seamlessly bridge image regions that are inconsistent with the projective model. However, the warping change of APAP in the adjacent areas is assumed to be small. In fact, the depth of the adjacent areas may change dramatically, which may still exhibit parallax artifacts in the vicinity of the object boundaries. Lee *et al.* divide an image into superpixels and propose the warping residual vectors to distinguish feature points from different depth planes [23].

Seam-Driven Methods. Seam-driven image stitching methods are also influential. A seam-cutting loss for the homography is proposed to measure the discontinuity between the warped target image and the reference image in the work of Gao *et al.*[12]. The homography with minimum seam-cutting loss is selected to achieve the best stitching. Zhang *et al.* [47] introduced content-

(a) The input of our deep image stitching framework: the reference image I_A and the target image I_B.

(b) The output of proposed large-baseline deep homogrpahy module: warped reference image I_{AW} and warped target image I_{BW}.

(c) Misalignments in overlapping areas caused by parallax.

(d) Removing artifacts at the cost of edge discontinuity.

(e) The output of proposed edge-preserved deformation module: the stitched image with edge continuity correction.

Fig. 1. The illustration of proposed edge-preserved image stitching strategy. (a)(b) demonstrate the input and output of the large-baseline homography module, learning to align the large-baseline inputs coarsely. (c)(d)(e) exhibit the effect of the edge-preserved deformation module, learning to eliminate the artifacts and smooth the discontinuous edges simultaneously.

preserving warping (CPW) [31] to align overlapping regions for small local adjustment while using the homography to maintain the global image structure. Different from aligning pixels of the overlapping area, Lin *et al.* [28] proposed to find a local area to stitch images, which can protect the curves and lines during stitching.

Although traditional image stitching methods have achieved promising performance, they cannot handle low-texture scenarios.

2.2. Deep image stitching

Deep image stitching is still in development, since the labeled data is hard to collect. In [37,50], synthetic datasets are proposed to solve this problem. Besides, a content revision network is proposed to generate the stitched image after image registration in [37].

However, the performance of these methods in real-world datasets is not reliable and the resolution of the network input is limited.

2.3. Deep Homography schemes

Homography estimation is an important part of image stitching, and deep homography can also be regarded as a significant step in deep image stitching. The deep homography solution was first proposed in [8], where a synthetic dataset and a VGG-style solution are put forward together. Then, Nguyen *et al.* [36] proposed an unsupervised version for [8], in which a photometric loss is adopted to measure the pixel error between warped images. Le *et al.* [22] and Zhang *et al.* [48] proposed content-aware networks to reject parallax regions and dynamic areas. And deep Lucas-Kanade networks [3,51] are also presented to align a template image with a source image. Besides, Koguciuk *et al.* [20] propose to increase the robustness using perceptual loss. Ye *et al.* [45] replace homography offset with motion basis to enhance the estimation performance.

Nevertheless, when it comes to scenes of low overlap rates, The performance of these solutions drops because of the limited receptive fields of convolutional layers.

3. Our method

In this section, we discuss our multi-scale deep homography module, edge-preserved deformation module, and size-free schemes, respectively.

3.1. Multi-scale deep homography

Although deep homography methods in scenes of high overlap rates [8,36,48,22,3] have outperformed traditional solutions, deep homography estimation in scenes of low overlap rates is still challenging due to the limited receptive fields of neural networks. To overcome this challenge, the proposed multi-scale deep homography network integrates feature pyramid and feature correlation into a network, increasing the utilization of feature maps and expanding the receptive field, respectively. The architecture of the proposed multi-scale deep homography network is shown in Fig. 2.

Feature Pyramid. After the images are fed into our network, they will be processed by 8 convolutional layers, where the number of filters per layer is set to 64, 64, 128, 128, 256, 256, 512, and 512, respectively. A max-pooling layer is adopted every two convolutional layers to represent multi-scale features as $F, F^{1/2}, F^{1/4}$, and $F^{1/8}$. As shown in Fig. 2, we select $F^{1/2}, F^{1/4}$, and $F^{1/8}$ to form a three-layer feature pyramid. The features of each layer in the pyramid are used to estimate the homography, and we transmit the estimated homography of the upper layer to the lower layer to enhance the prediction accuracy progressively. Besides, among the features of the four scales, the features of three scales will be used for subsequent homography regression, significantly improving the feature utilization.

Feature Correlation. To increase the receptive fields of our network, the feature correlation layer [38,14,39,18] is used here to strengthen feature matching explicitly. Formally, the correlation c between the reference feature $F_A^l \in W^l \times H^l \times C^l$ and the target feature $F_B^l \in W^l \times H^l \times C^l$ can be calculated as,

$$c(x_A^l, x_B^l) = \frac{< F_A^l(x_A^l), F_B^l(x_B^l) >}{|F_A^l(x_A^l)||F_B^l(x_B^l)|}, \quad x_A^l, x_B^l \in \mathbb{Z}^2, \tag{1}$$

where x_A^l, x_B^l are the 2-D spatial location in F_A^l and F_B^l, respectively. Specifying the search radius on the axis of width (or height) as R_w (or R_h), we obtain $c \in W^l \times H^l \times (2R_w + 1)\ (2R_h + 1)$ [43] by Eq. 1. Specifically, we calculate the global correlation by setting R_w (or R_h) equal to W^l (or H^l), and we calculate the local correlation when R_w (or R_h) is less than W^l (or H^l). By applying global correlation and local correlation to our network, we predict the homography progressively from global to local.

After extracting pyramid features and calculating feature correlations, we adopt a simple regression network that comprises three convolutional layers and two fully connected layers to predict eight vertices' offsets of the target image that can uniquely determine a homography. To be more specific, every layer of our three-layer pyramid predicts the residual offsets $\Delta_i, i = 1, 2, 3$. Every feature correlation in the pyramid is only calculated between the warped target feature and the reference feature rather than between the target feature and the reference feature. In this way, each layer in the pyramid only learns to predict the residual homography offsets instead of the complete offsets. And Δ_i can be calculated as follows:

$$\Delta_i = \mathscr{H}_{4pt}\left\{F_A^{1/2^{4-i}}, \mathscr{W}\left\langle F_B^{1/2^{4-i}}, \mathscr{DLT}\left(\sum_{n=0}^{i-1}\Delta_n\right)\right\rangle\right\}, \tag{2}$$

where $\mathscr{H}_{4pt}$ is the operation of estimating the residual offsets from the reference feature map and the warped target feature map. $\mathscr{W}$ warps the target feature map using the homography and $\mathscr{DLT}$ converts the offsets to the corresponding homography. We specify $\Delta_0 = 0$, which means all predicted offsets are 0. The final predicted offsets can be calculated as follows:

$$\Delta_{w\times h} = \Delta_1 + \Delta_2 + \Delta_3. \tag{3}$$

After that, image registration can be implemented by solving the homography and warping the input images.

Objective Function: Our multi-scale deep homography is trained in a supervised manner. Given the ground truth offsets $\hat{\Delta_{w\times h}}$, we designed the following objective function:

$$\begin{aligned}\mathscr{L}_H &= w_1\left(\hat{\Delta_{w\times h}} - \Delta_1\right)\\ &+ w_2\left(\hat{\Delta_{w\times h}} - \Delta_1 - \Delta_2\right)\\ &+ w_3\left(\hat{\Delta_{w\times h}} - \Delta_1 - \Delta_2 - \Delta_3\right),\end{aligned} \tag{4}$$

where the w_1, w_2, and w_3 represent the weights of each layer in the three-layer pyramid.

3.2. Edge-preserved deformation network

Stitching images with a global homography can easily produce artifacts in scenes with parallax. To eliminate the ghosting effects,

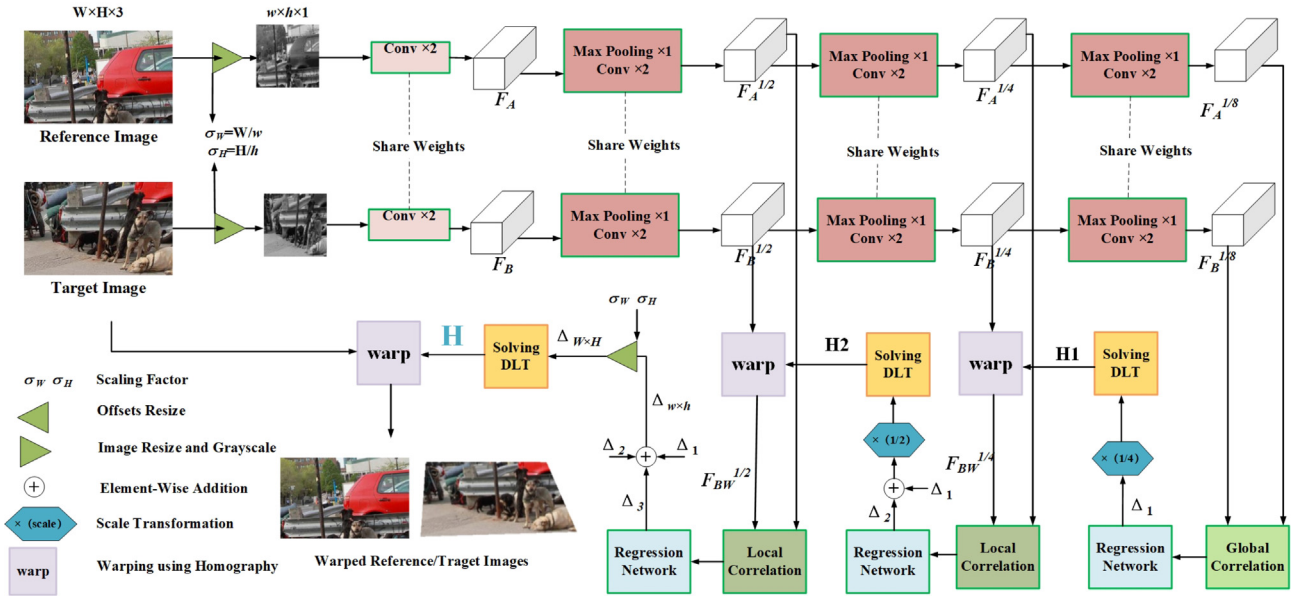

Fig. 2. The architecture of our multi-scale deep homography network.

we design an edge-preserved deformation network to learn the deformation rules of image stitching from edge to content. The learning process is quite different from traditional image fusion. As illustrated in Fig. 1(d)(e), this learning method first eliminates all the artifacts at the cost of edge discontinuity and then learns to correct the discontinuity at the strategy of edge-preserved.

Edge Deformation Branch. Compared with the rich information in an RGB image, such as color, texture, and content, the edge only contains the objects' contours in the image. Therefore, stitching the edges may be easier to achieve than stitching the RGB image. Inspired by this observation, we design an efficient approach to extract edges, and an edge deformation branch is used to stitch them. The edge map E for a grayscale image G can be obtained by calculating the difference of adjacent pixels as follows,

$$E_{i,j} = |G_{i,j} - G_{i-1,j}| + |G_{i,j} - G_{i,j-1}|, \tag{5}$$

where i and j are the horizontal and vertical coordinates. A convolutional layer with fixed kernels can achieve the operation to extract edges. Finally, we clip $E_{i,j}$ between 0 and 1. As for the edge deformation branch, we implement it using an encoder-decoder architecture as shown in Fig. 3 (middle). In this branch, the max-pooling or deconvolution is adopted every two convolutional layers and the number of convolutional kernels is set to 64, 64, 128, 128, 256, 256, 512, 512, 256, 256, 128, 128, 64, 64, and 1, respectively. Among these convolutional layers, the size of all kernels is set to 3×3 and the activation function is set to ReLU, except for the last convolutional layer. In the last layer, we set the kernel size to 1×1 and the activation function as Sigmoid to generate the stitched edge. Furthermore, to prevent the gradient vanishing problem and information imbalance in the training [40], skip connections are adopted to connect the low-level and high-level features with the same resolution.

Image Deformation Branch. We also design an image deformation branch to generate the stitched image in the guidance of the stitched edges. The image deformation branch has a similar architecture to the edge deformation branch as shown in Fig. 3 (top). To enable the image deformation branch of the edge-preserved stitching, we use the edge features learned by the edge deformation branch in the decoder stage to guide the learning. To be specific, we concatenate each feature map obtained by deconvolution in the edge deformation branch with the corresponding feature map in the image deformation branch from low-level to high-level. Besides, a fusion block is designed to integrate the last feature map in the edge deformation branch with the corresponding feature map in the image deformation branch, as illustrated in Fig. 4.

Objective Function. Similar to our deep homography, we train our stitching network in a supervised manner. To make the stitched edge close to the ground truth edge $\widehat{E}$ that is extracted from the ground truth image $\widehat{I}$, $\mathscr{L}_1$ loss is adopted as follows:

$$\mathscr{L}_{edge} = \frac{1}{W \times H \times 1} \left\| \widehat{E} - E \right\|_1, \tag{6}$$

where W and H define the width and height of the stitched edge.

Inspired by [15], we define a content loss to encourage our image deformation branch to generate perceptual naturally stitched images. Specifically, we use the 9-th convolutional layer in VGG-19 [42] as the representation of the image content. Let Φ_j denotes the j-th layer of VGG-19 and we define our content loss as follows:

$$\mathscr{L}_{content} = \frac{1}{W_j \times H_j \times C_j} \left\| \Phi_j\left(\widehat{I}\right) - \Phi_j(I) \right\|_2^2, \tag{7}$$

where W_j ,H_j and C_j denote the width, height, and channel number of the feature map, respectively.

Considering the constraints on the edge and content, we finally conclude our objective function as follows:

$$\mathscr{L}_S = \lambda_e \mathscr{L}_{edge} + \lambda_c \mathscr{L}_{content}, \tag{8}$$

where the λ_e and λ_c represent the balance factors of edge loss and content loss, respectively.

3.3. Size-free stitching

Size-free image stitching can be easily achieved by replacing the fully connected layers with convolutional layers [33]. However, the increase in input images' resolution will significantly increase the memory consumption because of feature correlation layers. Taking the global correlation as an example, the required memory can be

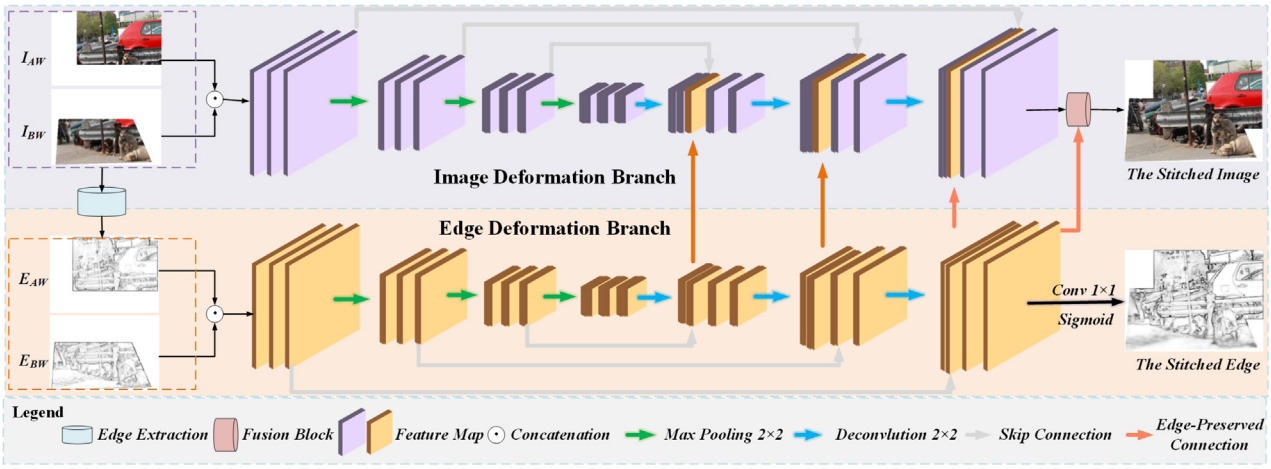

Fig. 3. The architecture of edge-preserved deformation network. Top: Image deformation branch. Middle: Edge deformation branch. Bottom: Legend.

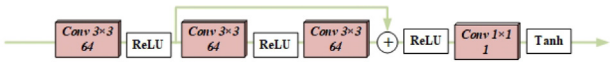

Fig. 4. The detail of the fusion block in the edge-preserved stitching branch.

expanded by λ^4 times when the resolution of input images is expanded by λ times. To make it more clear, we show the change of memory consumption as follows:

$$W^l \times H^l \times \left(2W^l + 1\right)\left(2H^l + 1\right) \Rightarrow \lambda W^l \times \lambda H^l \times \left(2\lambda W^l + 1\right)\left(2\lambda H^l + 1\right). \tag{9}$$

Obviously, adopting a fully convolutional network (FCN) cannot solve this problem. To reduce endless memory consumption, we design an alternative to achieve size-free stitching.

When we resize the images, we can change the corresponding offsets following the rule shown in Fig. 5. Noticing the relationship between image resize and offsets resize, we implement our size-free image stitching in three steps, as shown in Fig. 2: 1) We resize the input images from $W \times H$ to $w \times h$ and save scaling factors for width and height σ_W, σ_H. 2) We predict the offsets from the images of $w \times h$. 3) We resize the offsets using σ_W and σ_H by the rule shown in Fig. 5 to make them correspond to the images of $W \times H$. In short, we complete size-free homography estimation using the relationship between image resize and offsets resize without extra memory consumption. Since the edge-preserved deformation module can be regarded as an FCN, our deep image stitching framework can process arbitrary size inputs.

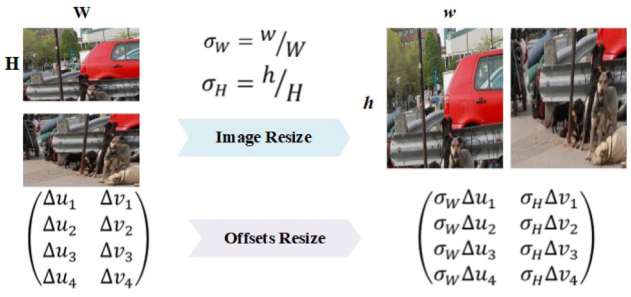

Fig. 5. The relationship between image resize and offsets resize. $(\Delta u_i, \Delta v_i)$ represents the coordinate offsets of the i-th vertex in the target image, where i = 1, 2, 3, and 4.

4. Experiments

In this section, we carry out experiments to validate the effectiveness of our method.

4.1. Dataset and implementation details

Dataset. Deep homography and deep image stitching are two different tasks, but we adopt the same dataset to train them together. We follow the strategy of [37] to generate a seemingly infinite dataset for image stitching from Microsoft COCO [29]. We call this dataset Stitched MS-COCO, and we demonstrate some samples in Fig. 6. To be specific, in addition to the random perturbation $[-\rho, \rho]$ [8] of the four vertices in an image patch, the random translation $[-\tau, \tau]$ [37] is added to simulate the characteristics of low overlap rates in image stitching. The format of Stitched MS-COCO can be described as a quadruple $(I_{Reference}, I_{Target}, \Delta, Label)$, of which $I_{Reference}$ and I_{Target} represent the reference image and target image to be stitched, Δ represents the 8 coordinate offsets of the four vertices to estimate a homography, and *Label* is the ground truth of the stitched result. Specifically, when generating a quadruple from a real image $(W \times H)$, we set the size of image patches $(P^W \times P^H)$ to be input into our network to $W/2.4 \times H/2.4$, the maximum translation $(\tau^W \times \tau^H)$ to $0.5P^W \times 0.5P^H$, and the maximum perturbance $(\rho^W \times \rho^H)$ to $0.2P^W \times 0.2P^H$. Moreover, Δ can be calculated by adding translation and perturbance together. We generate 50,000 quadruples from MS-COCO train2014 as the training set and 5,000 quadruples from test2014 as the testing set.

Details. The training process is completed in two steps: deep homography module and deep deformation module. Our deep

Fig. 6. Several samples of our Stitched MS-COCO dataset. Each sample is separated by a dashed line. The $I_{Reference}, I_{Target}$ and *Label* are demonstrated in each instance.

homography network is trained by an Adam optimizer [19] for up to 100 epochs, with an exponentially decaying learning rate initialized as 10^{-4}, a decay step of 12,500, and a decay rate of 0.95. According to the different influence of each pyramid layer on the homography prediction, we set w_1, w_2, and w_3 to 1, 0.25, and 0.1, respectively. We adopt some data augmentation techniques to enhance illumination robustness, such as artificially inserting random brightness shifts into the training images. Subsequently, we train our stitching module with the parameters of the homography network being fixed. The training strategy is the same as that of the homography module, except for the maximum training epoch set as 25. The balance factors λ_e and λ_c are set to 1 and $2e^{-6}$. In addition, the batch size numbers of the two training steps are set to 4 and 1. The input size $W \times H$ of our framework is arbitrary, and the scaling size $w \times h$ is set to 128×128 which is consistent with [8,36,48,22]. All the components of this framework are implemented on TensorFlow, and the training process is performed on one NVIDIA RTX 2080 Ti.

4.2. Comparison with homography estimations

Traditional homography estimations differ according to different feature descriptors and different outlier rejections. In our experiments, we choose SIFT [35] and ORB [41] as the feature descriptors. RANSAC [10] and MAGSAC [2] are chosen as the outlier rejection algorithms. Besides that, we compare our method with deep homography algorithms, including DHN [8], UDHN [36], and CA-UDHN [48]. Since the labeled homography can be obtained in the synthetic datasets, we adopt the 4pt-Homography RMSE as the evaluation metric, which is also used in [36].

Warped MS-COCO. Warped MS-COCO, which only includes the random perturbance $[-\rho, \rho]$ of four vertices, is the most widely acknowledged synthetic dataset for deep homography estimation. We first conduct a comparative experiment on this dataset with $\rho = 32$, where each corner of the image patch can be perturbed by a maximum of one-quarter of the total image size. The results are shown in Table 1, where $I_{3\times3}$ refers to a 3×3 identity matrix as a 'no-warping' homography for reference. The performance of traditional homography solutions heavily relies on the quality of hand-craft feature points, which indicates this method may fail in low-texture scenes. To avoid this problem, we set the estimated homography to the identity matrix when that happens. As shown in Table 1, the results are divided into three parts to illustrate each method's various performance profiles as follows:

(1) The four traditional methods perform pretty well in the top 60% of all the testing sets, while it usually cannot capture enough matched features to estimate a homography in the worst 40% of all.
(2) UDHN and DHN achieve similar performance with offsets' errors controlled to several pixels all the time.
(3) CA-UDHN achieves state-of-the-art performance in small-baseline scenes, while its performance is close to $I_{3\times3}$ in case of low overlap rates. This result is due to its limited perception field, thus making it unable to perceive the long-range matching information between the two inputs.
(4) Our multi-scale deep homography solution outperforms all the compared deep solutions and traditional methods with a large margin all the time.

Stitched MS-COCO. In most cases of image stitching, the overlap rate between images is much lower than that in Warped MS-COCO. Here, the existing homography estimation solutions' performance drops sharply as the overlap rate decreases, while our method is still robust and accurate. We verified this view on Stitched MS-COCO dataset that is much more challenging due to the larger displacement and the lower overlap rates. To be consistent with Warped MS-COCO, we resize $I_{Reference}$ and I_{Target} to 128×128 in this experiment. Compared with the supervised solution DHN, the unsupervised solution UDHN requires extra information around the image patch to prevent ambiguity during the training process [36,48]. However, Stitched MS-COCO is only composed of image patches and corresponding homography offsets, making UDHN unable to be trained on this dataset. Therefore, we test UDHN using the model trained on Warped MS-COCO.

In addition to these methods, we also compare ours with GC-DHN [37] — the deep homography network of the first deep image stitching method. The results are shown in Fig. 7 and we can conclude that:

(1) As the overlap rate decreases, the accuracy of all methods continues to decrease, of which the accuracy of SIFT + RANSAC, DHN, and UDHN decreases faster than other methods.
(2) The lower the overlap rate is, the closer the performance of the SIFT + RANSAC, DHN, and UDHN is to $I_{3\times3}$, which indicates that these methods may fail to work when the overlap rate is particularly low.
(3) Our solution outperforms the deep homography network (GC-DHN) in other deep image stitching work [37].
(4) Our solution can maintain good accuracy even at low overlap rates. This benefits from the combination of feature pyramid and feature correlation, which explicitly increase the network's receptive field on the feature maps of different scales.

4.3. Comparison with image stitching algorithms

Deep image stitching algorithms are still in development, we choose VFISNet [37], a view-free image stitching network, as a rep-

Table 1
Comparison experiment for homography estimation on Warped MS-COCO ($\rho = 32$). The number represents the 4pt-Homography RMSE between the estimated offsets of 4 vertexs and the ground truth. All the learning methods are trained on Warped MS-COCO. *F* indicates that this method is worse than $I_{3\times3}$ in the current dataset.

Method	Top 0∼30%	30∼60%	60∼100%	Average
$I_{3\times3}$	15.0154	18.2515	21.3548	18.5220
SIFT[35]+RANSAC[10]	0.6687	1.1223	18.5990	7.9769
SIFT[35]+MAGSAC[2]	0.5697	0.8679	*F*	*F*
ORB[41]+RANSAC[10]	3.8995	10.2206	*F*	*F*
ORB[41]+MAGSAC[2]	3.1557	8.7443	*F*	*F*
DHN[8]	3.2998	4.8839	7.7017	5.5358
UDHN[36]	2.1894	3.5272	6.5073	4.3179
CA-UDHN[48]	15.0082	18.2498	*F*	*F*
Ours	**0.2719**	**0.4140**	**0.9761**	**0.5962**

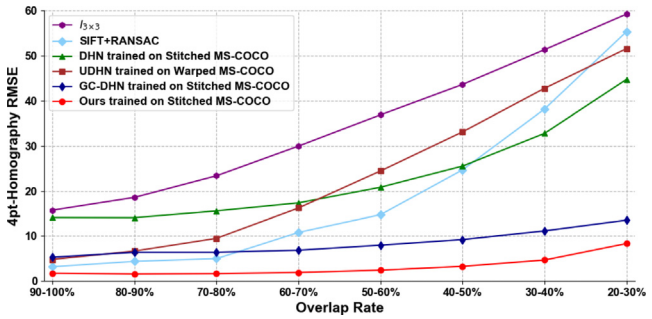

Fig. 7. Comparative experiment for homography estimation on Stitched MS-COCO ($\tau = 64, \rho = 25$).

Table 2
Quantitative comparison between VFISNet and ours on Stitched MS-COCO.

Method	PSNR	SSIM
VFISNet [37]	24.8525	0.9241
Ours	**27.4462**	**0.9463**

resentative of deep image stitching to compare. Since its input size is 128×128, we combine it with Bicubic interpolation to produce the stitched results of arbitrary size. As for traditional methods, we compare our method with four classical image stitching algorithms: Global Homography, SPHP [5], APAP[46], and robust ELA [25], in which the first two are classic methods with global transformation models and the others are with local adaptive stitching fields. Among these four methods, we implement Global Homography using SIFT, RANSAC, and average fusion. The results of SPHP, APAP, and robust ELA are obtained by running their open-source codes with our testing instances. These methods are evaluated on our synthetic images and real images, respectively.

Synthetic Images. First of all, we conduct a quantitative comparison between VFISNet and ours as shown in Table 2. Since the two deep solutions adopt the same dataset, it is easy to calculate the PSNR and SSIM. As for the traditional methods, the resolution of the stitched images differs according to different ways, it's hard to compare them with ours quantitatively.

And a qualitative comparison is carried on in our synthetic dataset in Fig. 8. There are apparent artifacts in the stitched result of Global Homography because the mismatch of feature points affects homography estimation accuracy. Compared with SPHP, APAP, and robust ELA, our solution shows competitive performance with these classic and convincing image stitching works. In deep image stitching methods, our results are visually more clear than that of VFISNet + Bicubic.

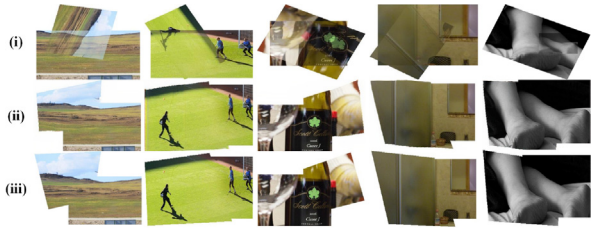

Fig. 9. Failure cases of traditional feature-based methods. (i)(ii): The stitched results of the Global Homography and ours. (iii): The ground truth.

Besides that, our method is more robust. Traditional methods heavily depend on the quality of feature detection and feature matching. However, the feature points can be easily affected by various environments. We test 1,000 pairs of images in our test set with the Global Homography and our method. Experimental results show that more than 30 pairs fail using the Global Homography, while all work in our method. Fig. 9 shows some failure cases of traditional methods in our synthetic dataset. As for other feature-based methods, the number of failures can be much more than that of the Global Homography, because they usually have stricter requirements on the distribution or number of feature points. For instance, APAP would require more feature points to find a valid point subset when generating hypotheses for multi-structure data [7]. The robustness of our method benefits from the robust deep features that are adaptively learned in a neural network.

Real Images. In addition to synthetic images, we also conduct a cross-dataset experiment, in which our model is tested on real images with varying degrees of parallax. Although our method is only trained on a synthetic dataset without parallax, the proposed edge-preserved deformation module enables our deep framework the ability to handle misalignments caused by parallax.

As shown in Fig. 10, the first 5 examples come from classic image stitching cases that are widely used in existing traditional image stitching methods, and the last 5 are challenging cases with obvious parallax or even moving objects taken by ourselves. The arrows highlight the artifacts. Due to GPU memory limitation, we limit the input images' maximum size not to exceed 512×512. From the results shown in Fig. 10, we can observe:

(1) The deep image stitching methods (VFISNet and ours) can eliminate almost all the artifacts, while the traditional methods (Global Homography, SPHP, APAP, robust ELA) cannot do it in various stitching scenes. This phenomenon can be attributed to different stitching strategies. To eliminate the artifacts, the traditional solutions try to align the reference image and target image as much as possible. However, the stitching quality heav-

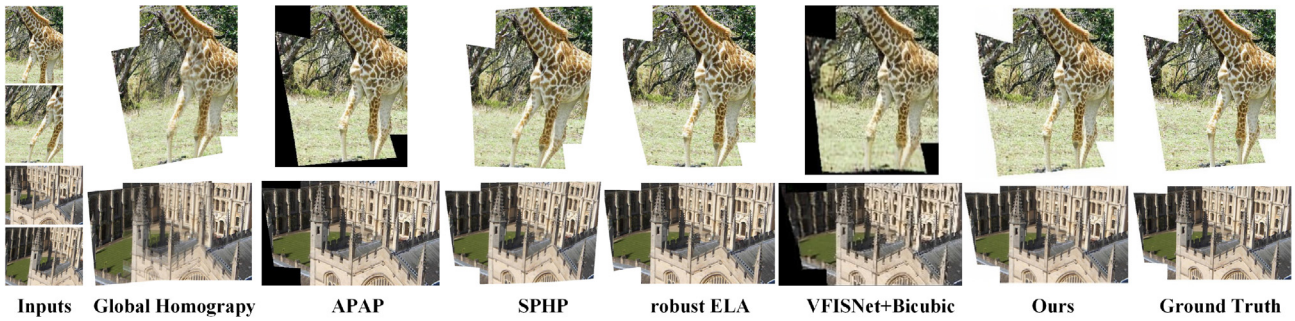

Fig. 8. The comparative experiments in our synthetic dataset. Col 1: Input images. Col 2–7: Stitched results of the global homography, SPHP [5], APAP [46], robust ELA [25], VFISNet[37]+Bicubic, and ours. Col 8: T.he ground truth.

ily relies on the number and distribution of the feature points, failing to eliminate the ghosting effects in varying scenes. As for the proposed deep image stitching, the network tends to learn the overlapping areas from the reference image, neglecting the target image and free from the artifacts. Although this learning tendency may make the edges discontinuous, our network would learn to revise it to look smooth and natural.

(2) Our method outperforms the existing deep image stitching method. Although the deep solutions can eliminate the artifacts, they bring another problem: the stitched images' non-overlapping regions are blurred and discontinuous. This problem can be observed obviously in the results of VFISNet + Bicubic, while our method alleviates this problem by learning image stitching from edge to content progressively.

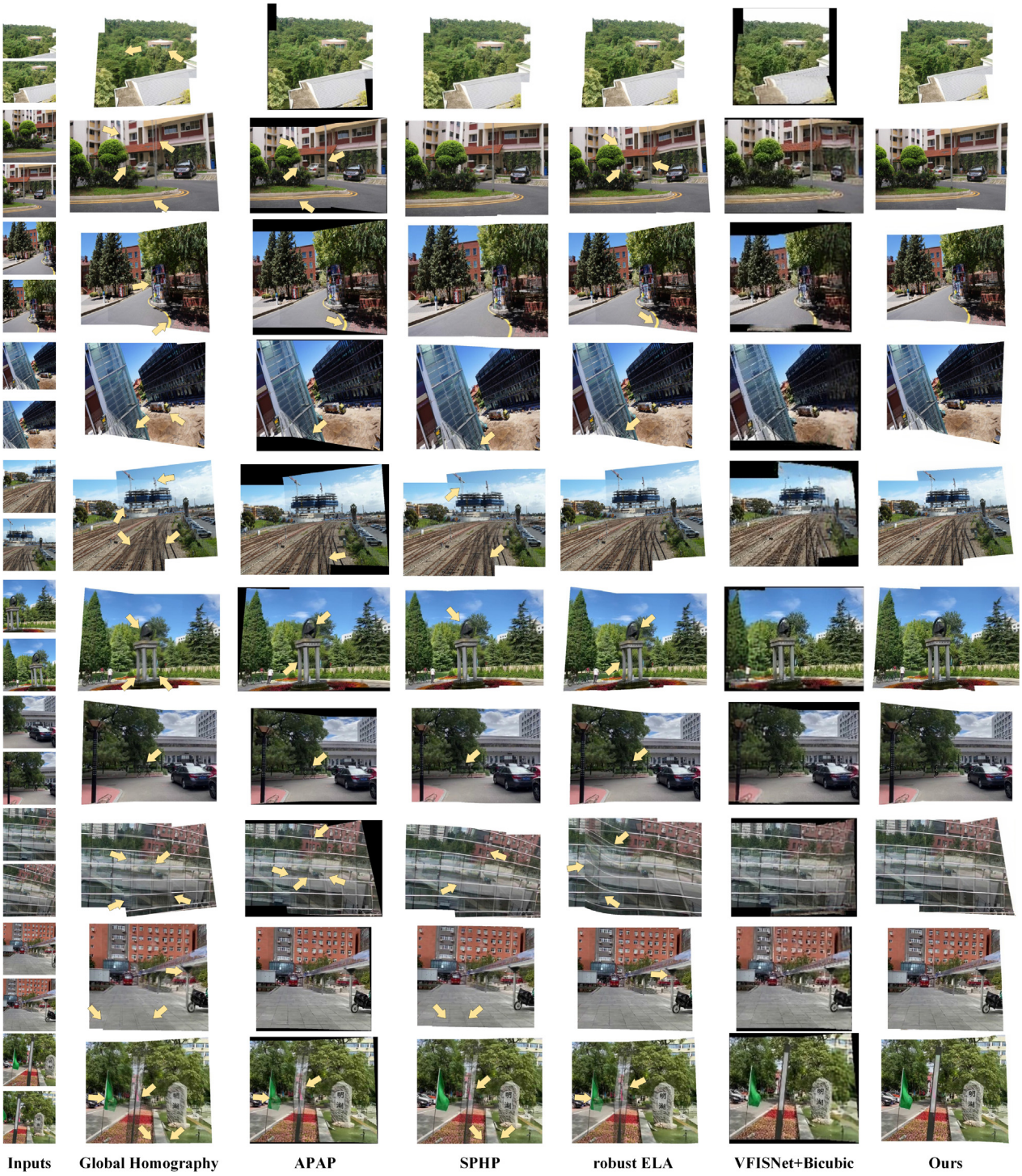

Fig. 10. The comparative experiments in real images. Col 1: Input images. Col 2–7: Stitching results of the global homography, SPHP [5], APAP [46], robust ELA [25], VFISNet [37]+Bicubic, and ours. The first 5 examples come from classic image stitching cases (“roof” [49], “yard” [12], “site” [46], “construction” [46] and “railtrack” [46]), and the last 5 are challenging cases with obvious parallax or even moving person tak.en by ourselves.

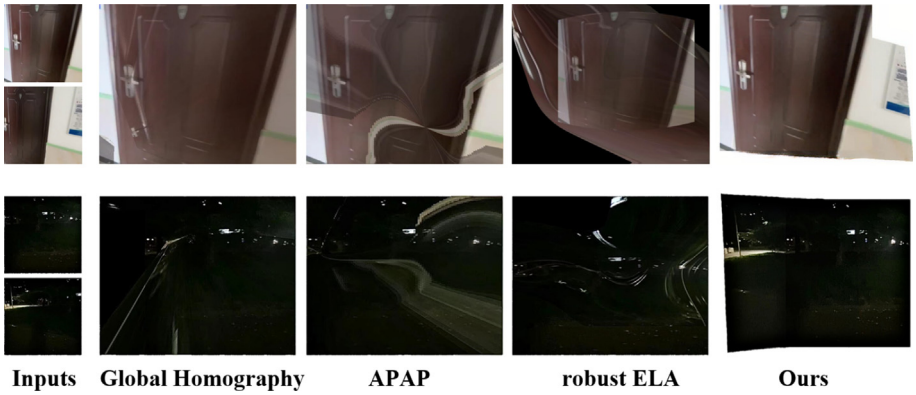

Fig. 11. Robustness comparison in real low-texture scenes. Row 1: An indoor scene. Row 2: A low-light scene.

(3) In a scene containing moving objects, the learning methods perform better than the traditional methods. Row 7 of Fig. 10 exhibits a pair of images that contains a moving person. We can see that the Global Homography, SPHP, APAP, and robust ELA cannot handle this moving person while the learning methods deal with it successfully.

In addition, we compare the robustness in real challenging scenes. As shown in Fig. 11, traditional solutions fail due to the poor quality of hand-craft feature points in low-texture scenes, while the proposed deep solution succeeds because of the learnable robust deep features.

4.4. Ablation studies

We conduct ablation experiments to validate the necessity of each part in our proposed framework.

Feature Pyramid. The feature pyramid serves as a multi-scale feature extractor in our method. To reduce parameters, we set the kernel size of each convolutional layer to 3×3. However, the receptive field of the 3×3 kernel is significantly limited. To mitigate this contradiction, the feature pyramid is adopted to extract multi-scale features on different pyramid levels with a fixed kernel size. We evaluate the significance of the feature pyramid with our synthetic dataset on the homography estimation task. As we can see in Fig. 12, our complete pyramid model has significantly smaller errors than one-layer or two-layer models.

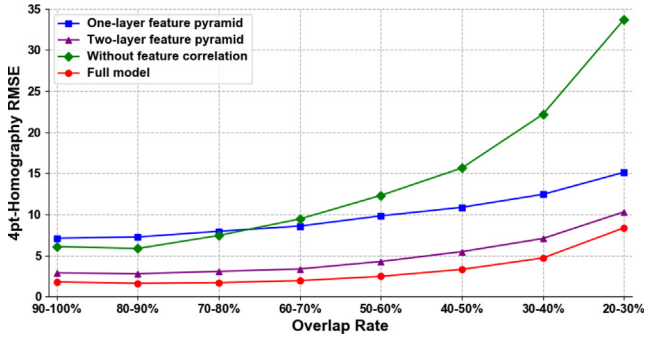

Fig. 12. Ablation experiments on feature pyramid and feature correlation for homography estimation. Feature pyramid: The three-layer pyramid model is better than one-layer and two-layer. Feature correlation: The model with feature correlation is better than that without.

Feature Correlation. The feature correlation layer plays the role of feature matching in our method. Different from other deep homography estimations [8,36,48,22] that match features by learning convolutional filters, our feature correlation layers match features by making full use of the features extracted by the convolutional layers. Besides that, our global-to-local strategy ensures our capability to match features all over feature maps. To validate the effects of feature correlation, we experiment with removing feature correlation layers, where both the global correlation and the local correlation are ablated. The results are shown in Fig. 12, where the RMSE increases with a large margin in the absence of feature correlation, especially with the low overlap rate.

Edge Deformation Branch. To validate the effectiveness of the edge deformation branch, we carry out ablation experiments on real images. We retrain the deformation module without the edge deformation branch. The results are illustrated in Fig. 13, and we can observe:

(1) With or without the edge deformation branch, the network can learn to eliminate artifacts in the overlapping area.
(2) After ablating this branch, the edges of the stitched images are not discontinuous as shown in Fig. 13 (a). With this branch (Fig. 13 (b)), the network further learns to smooth the discontinuous edges, contributing to visually pleasing and edge-continuity stitched results.

5. Conclusion

This paper presents a novel deep image stitching algorithm that can stitch images from arbitrary shooting positions into a perceptually natural image. First, a multi-scale deep homography network is proposed to implement homography estimation and image registration, which outperforms existing deep solutions and traditional solutions with a large margin. Then we present an edge-preserved deformation module to learn the deformation rules of image stitching from the warped images. Furthermore, some schemes are adopted to enable our network the capability of free-size stitching when the fully connected layers are inevitable. Experiments show that our superiority to the existing learning method and shows competitive performance with state-of-the-art traditional methods. Furthermore, as a learning method that is only trained in a synthetic dataset, our method exhibits excellent generalization in other real-world datasets.

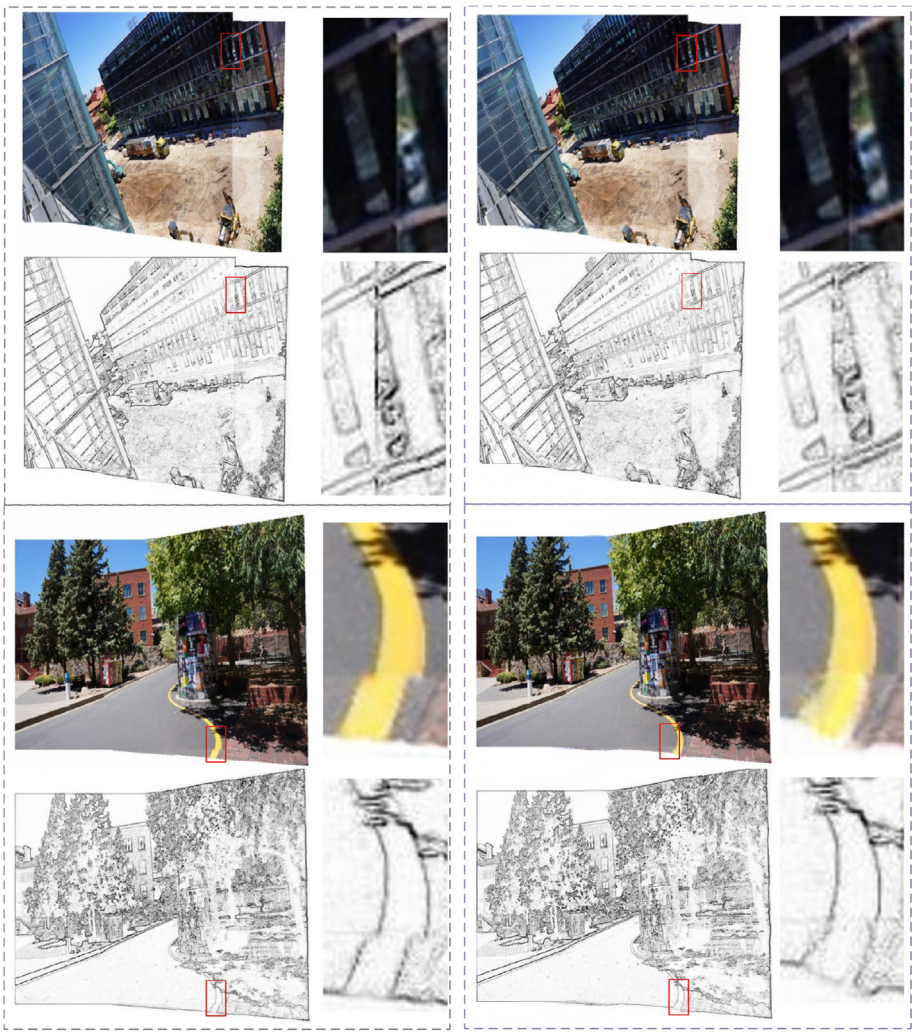

Fig. 13. Ablation experiment on real images to validate the effects of edge deformation branch.

CRediT authorship contribution statement

Lang Nie: Conceptualization, Methodology, Software. **Chunyu Lin:** Methodology, Supervision. **Kang Liao:** Visualization, Investigation. **Yao Zhao:** Supervision.

Declaration of Competing 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.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Nos. 62172032, 61772066, 61972028).

References

[1] R. Anderson, D. Gallup, J.T. Barron, J. Kontkanen, N. Snavely, C. Hernández, S. Agarwal, S.M. Seitz, Jump: virtual reality video, ACM Trans. Graph. (TOG) 35 (2016) 1–13.
[2] D. Barath, J. Matas, J. Noskova, Magsac: marginalizing sample consensus, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10197–10205.
[3] C.H. Chang, C.N. Chou, E.Y. Chang, Clkn: Cascaded lucas-kanade networks for image alignment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2213–2221.
[4] C.H. Chang, Y.Y. Chuang, A line-structure-preserving approach to image resizing, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition IEEE, 2012, pp. 1075–1082.
[5] C.H. Chang, Y. Sato, Y.Y. Chuang, Shape-preserving half-projective warps for image stitching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3254–3261.
[6] Y.S. Chen, Y.Y. Chuang, Natural image stitching with the global similarity prior, European conference on computer vision, Springer (2016) 186–201.
[7] T.J. Chin, J. Yu, D. Suter, Accelerated hypothesis generation for multistructure data via preference analysis, IEEE Trans. Pattern Anal. Mach. Intell. 34 (2011) 625–638.
[8] D. DeTone, T. Malisiewicz, A. Rabinovich, Deep image homography estimation. arXiv preprint arXiv:1606.03798, 2016..
[9] A. Eden, M. Uyttendaele, R. Szeliski, Seamless image stitching of scenes with large motions and exposure differences, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), IEEE, 2006. pp. 2498–2505..
[10] M.A. Fischler, R.C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Commun. ACM 24 (1981) 381–395.
[11] J. Gao, S.J. Kim, M.S. Brown, Constructing image panoramas using dual-homography warping, in: CVPR 2011, IEEE, 2011. pp. 49–56..
[12] J. Gao, Y. Li, T.J. Chin, M.S. Brown, Seam-driven image stitching., in: Eurographics (Short Papers), 2013, pp. 45–48..
[13] R. Hartley, A. Zisserman, Multiple view geometry in computer vision, Cambridge University Press, 2003.
[14] S. Huang, Q. Wang, S. Zhang, S. Yan, X. He, Dynamic context correspondence network for semantic alignment, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2010–2019.

[15] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, European conference on computer vision, Springer (2016) 694–711.
[16] S. Kasahara, S. Nagai, J. Rekimoto, Jackin head: Immersive visual telepresence system with omnidirectional wearable camera, IEEE Trans. Visualiz. Comput. Graph. 23 (2016) 1222–1234.
[17] H.G. Kim, H.T. Lim, Y.M. Ro, Deep virtual reality image quality assessment with human perception guider for omnidirectional image, IEEE Trans. Circuits Syst. Video Technol. 30 (2019) 917–928.
[18] S. Kim, S. Lin, S. Jeon, D. Min, K. Sohn, Recurrent transformer networks for semantic correspondence, 2018. arXiv preprint arXiv:1810.12155..
[19] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014..
[20] D. Koguciuk, E. Arani, B. Zonooz, Perceptual loss for robust unsupervised homography estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4274–4283.
[21] W.S. Lai, O. Gallo, J. Gu, D. Sun, M.H. Yang, J. Kautz, Video stitching for linear camera arrays, 2019. arXiv preprint arXiv:1907.13622..
[22] H. Le, F. Liu, S. Zhang, A. Agarwala, Deep homography estimation for dynamic scenes, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7652–7661.
[23] K.Y. Lee, J.Y. Sim, Warping residual based image stitching for large parallax, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8198–8206.
[24] J. Li, B. Deng, R. Tang, Z. Wang, Y. Yan, Local-adaptive image alignment based on triangular facet approximation, IEEE Trans. Image Process. 29 (2019) 2356–2369.
[25] J. Li, Z. Wang, S. Lai, Y. Zhai, M. Zhang, Parallax-tolerant image stitching based on robust elastic warping, IEEE Trans. Multimedia 20 (2017) 1672–1687.
[26] N. Li, Y. Xu, C. Wang, Quasi-homography warps in image stitching, IEEE Trans. Multimedia 20 (2017) 1365–1375.
[27] C.C. Lin, S.U. Pankanti, K. Natesan Ramamurthy, A.Y. Aravkin, Adaptive as-natural-as-possible image stitching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1155–1163.
[28] K. Lin, N. Jiang, L.F. Cheong, M. Do, J. Lu, Seagull: Seam-guided local alignment for parallax-tolerant image stitching, European conference on computer vision, Springer (2016) 370–385.
[29] T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, European conference on computer vision, Springer (2014) 740–755.
[30] W.Y. Lin, S. Liu, Y. Matsushita, T.T. Ng, L.F. Cheong, Smoothly varying affine stitching, in: CVPR 2011, IEEE, 2011. pp. 345–352..
[31] F. Liu, M. Gleicher, H. Jin, A. Agarwala, Content-preserving warps for 3d video stabilization, ACM Trans. Graph. (TOG) 28 (2009) 1–9.
[32] S. Liu, Q. Chai, Shape-optimizing and illumination-smoothing image stitching, IEEE Trans. Multimedia 21 (2018) 690–703.
[33] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.
[34] Z. Lou, T. Gevers, Image alignment by piecewise planar region matching, IEEE Trans. Multimedia 16 (2014) 2052–2061.
[35] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2004) 91–110.
[36] T. Nguyen, S.W. Chen, S.S. Shivakumar, C.J. Taylor, V. Kumar, Unsupervised deep homography: a fast and robust homography estimation model, IEEE Robot. Autom. Lett. 3 (2018) 2346–2353.
[37] L. Nie, C. Lin, K. Liao, M. Liu, Y. Zhao, A view-free image stitching network based on global homography, J. Vis. Commun. Image Represent. 102950 (2020).
[38] I. Rocco, R. Arandjelović, J. Sivic, Efficient neighbourhood consensus networks via submanifold sparse convolutions, European Conference on Computer Vision, Springer. (2020) 605–621.
[39] I. Rocco, M. Cimpoi, R. Arandjelović, A. Torii, T. Pajdla, J. Sivic, Neighbourhood consensus networks, 2018. arXiv preprint arXiv:1810.10510..
[40] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, International Conference on Medical image computing and computer-assisted intervention, Springer. (2015) 234–241.
[41] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, Orb: An efficient alternative to sift or surf, in: 2011 International conference on computer vision Ieee, 2011, pp. 2564–2571.
[42] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014. arXiv preprint arXiv:1409.1556..
[43] D. Sun, X. Yang, M.Y. Liu, J. Kautz, Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8934–8943.
[44] L. Wang, W. Yu, B. Li, Multi-scenes image stitching based on autonomous driving, in: 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), IEEE, 2020, pp. 694–698.
[45] N. Ye, C. Wang, H. Fan, S. Liu, Motion basis learning for unsupervised deep homography estimation with subspace projection, 2021. arXiv preprint arXiv:2103.15346..
[46] J. Zaragoza, T.J. Chin, Q.H. Tran, M.S. Brown, D. Suter, As-projective-as-possible image stitching with moving dlt, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2014) 1285–1298, https://doi.org/10.1109/TPAMI.2013.247.
[47] F. Zhang, F. Liu, Parallax-tolerant image stitching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3262–3269..
[48] J. Zhang, C. Wang, S. Liu, L. Jia, N. Ye, J. Wang, J. Zhou, J. Sun, Content-aware unsupervised deep homography estimation, European Conference on Computer Vision, Springer. (2020) 653–669.
[49] Y. Zhang, Y.K. Lai, F.L. Zhang, Content-preserving image stitching with piecewise rectangular boundary constraints, IEEE Trans. Visual Comput. Graphics 27 (2021) 3198–3212, https://doi.org/10.1109/TVCG.2020.2965097.
[50] Q. Zhao, Y. Ma, C. Zhu, C. Yao, B. Feng, F. Dai, Image stitching via deep homography estimation, Neurocomputing 450 (2021) 219–229.
[51] Y. Zhao, X. Huang, Z. Zhang, Deep lucas-kanade homography for multimodal image alignment, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15950–15959.

Lang Nie received the B.S. degree in computer science and technology from Beijing Jiaotong University, Beijing, China, in 2019, where he is currently pursuing the Ph.D. degree in signal and information processing with the Institute of Information Science. His current research interests include image and video processing, 3D vision, and multi-view geometry.

Chunyu Lin received the Doctor degree from Beijing Jiaotong University (BJTU), Beijing, China, in 2011. From 2009 to 2010, he was a Visiting Researcher at the ICT Group, Delft University of Technology, Netherlands. From 2011 to 2012, he was a Post-Doctoral Researcher with the Multimedia Laboratory, Gent University, Belgium. He is currently a full professor in BJTU. His research interests include image/video compression and robust transmission, 3-D visual analysis, Vision-based Advanced Driver Assistance Systems.

Kang Liao received the B.S degree in software engineering from Shaanxi Normal University, Xi'an, Shaanxi, China, in 2017, and is currently pursuing the Ph.D. degree in signal and information processing from the Institute of Information Science, Beijing Jiaotong University (BJTU), Beijing, China. His current research interests include image and video processing, 3-D scene understanding, and adversarial learning.

Yao Zhao received the B.S. degree from Fuzhou University, Fuzhou, China, in 1989, and the M.E. degree from Southeast University, Nanjing, China, in 1992, both from the Radio Engineering Department, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), Beijing, China, in 1996. He became an Associate Professor at BJTU in 1998 and became a Professor in 2001. From 2001 to 2002, he was a Senior Research Fellow with the Information and Communication Theory Group, Faculty of Information Technology and Systems, Delft University of Technology, Delft, The Netherlands. He is currently the Director of the Institute of Information Science, BJTU. His current research interests include image/video coding, digital watermarking and forensics, and video analysis and understanding. He serves on the Editorial Boards of several international journals, including as an Associate Editor of the IEEE TRANSACTIONS ON CYBERNETICS, a Senior Associate Editor of the IEEE SIGNAL PROCESSING LETTERS, and an Area Editor of Signal Processing: Image Communication. He was named a Distinguished Young Scholar by the National Science Foundation of China in 2010, and was elected as a Chang Jiang Scholar of Ministry of Education of China in 2013. He is a Fellow of the IET.