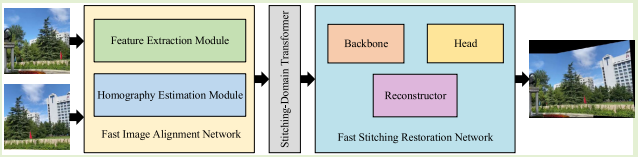
A Fast Unsupervised Image Stitching Model Based on Homography Estimation

Jianjun Ni, *Senior Member, IEEE*, Yingqi Li, Chunyan Ke, Ziru Zhang, Weidong Cao, *Member, IEEE*, and Simon X. Yang, *Senior Member, IEEE*

***Abstract*—Image stitching is the synthesis of multiple partial image segments into a complete and continuous panoramic image through effective image alignment and seamless fusion techniques. It can achieve a wider field of view and richer information for display and analysis. Most deep learning-based image stitching methods have significant advantages in improving accuracy, but they are not suitable for real-time applications due to multiple iterations of computation or deeper network depth. To deal with this problem, a fast unsupervised image stitching model is proposed in this article. In the proposed model, an adaptive feature extraction module (FEM) for deformation is designed, and then a fast unsupervised learning-based image alignment network is proposed. In addition, a stitching restoration network with a smaller number of parameters is presented to remove the redundant and unnecessary sampling and convolution operations in general deep learning-based models. Finally, some experiments are conducted on both the synthetic and real-scene datasets. The total stitching accuracy of the proposed model is higher, and the details of the output images are clearer. The proposed can achieve 1.79, 26.54, and 0.86 in RMSE, peak signal-to-noise ratio (PSNR), and structural similarity (SSIM) on the alignment results, respectively, which are better than those of the state-of-the-art methods. Furthermore, the comparison results prove that the proposed model can effectively reduce memory loss, and achieve a fast unsupervised image stitching, with a very small model size.**



***Index Terms*— Deformation invariance, fast alignment, homography estimation, image stitching.**

I. Introduction

IMAGE stitching is a technique widely used in computer vision and image processing. It is a process of combining multiple images with narrow but overlapping fields of view to create a larger image with a wider field of view [1], [2], [3]. Currently, it is widely used in various fields, such as map making [4], virtual reality [5], autonomous driving [6], [7], and medical image processing [8].

The design of image stitching models involves a lot of imaging processing methods, such as feature point detection, homography estimation, matching alignment, and image fusion [9], [10], [11], [12]. There are four main steps in the classical image stitching model, namely, image matching, reprojection, image stitching, and image fusion [13]. In these steps, image matching is a crucial step for image stitching. If the matching is inaccurate, it will lead to discontinuity and unnatural seams in the final stitched image.

Most of the traditional image stitching methods are based on underlying feature point matching [14], [15], such as the features obtained by scale-invariant feature transform (SIFT) [16], [17] and speeded-up robust features (SURFs) [18]. These methods rely heavily on the precise localization and uniform distribution of hand-crafted sparse features, which are very sensitive to the influence of environmental factors, such as lighting, viewing angle, and occlusion.

With the recent rapid development of deep learning [19], [20], [21], some authors have proposed deep neural network-based homography estimation methods [22], [23], [24], [25]. These deep learning-based methods have greatly improved the accuracy of image matching. However, most of these homography estimation methods are only for the case that the input image has with small displacement and large overlap, which is not reasonable for the image stitching problem. If the stitching is just achieved by directly fusing the results of homography estimation, the small estimation errors will cause bad visual effects.

Manuscript received 17 July 2024; accepted 27 July 2024. Date of publication 6 August 2024; date of current version 13 September 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 61873086, in part by the Jiangsu Province Key Research and Development Program under Grant BE2023340, and in part by the Changzhou Science and Technology Bureau Program under Grant CJ20230045. The associate editor coordinating the review of this article and approving it for publication was Dr. Xiaojin Zhao. *(Corresponding author: Jianjun Ni.)*

Jianjun Ni, Yingqi Li, Chunyan Ke, Ziru Zhang, and Weidong Cao are with the College of Artificial Intelligence and Automation, Hohai University, Changzhou, Jiangsu 213200, China (e-mail: njjhhuc@gmail.com; liyingqi@hhu.edu.cn; chunyanke@hhu.edu.cn; ziru_zhang@hhu.edu.cn; cwd2018@hhu.edu.cn).

Simon X. Yang is with the Advanced Robotics and Intelligent Systems (ARIS) Laboratory, School of Engineering, University of Guelph, Guelph, ON N1G 2W1, Canada (e-mail: syang@uoguelph.ca).

Digital Object Identifier 10.1109/JSEN.2024.3436051

Facing this problem, more and more researchers have proposed improved image stitching methods with different architectures, such as VFISNet [26], EPISNet [27], and so on [28], [29]. These models have obtained great research progress in the image stitching field. However, all of these solutions are based on supervised deep learning methods. In real scenes, the existing images do not give us stitching labels, and there is no image stitching dataset for real scenes with supervised learning methods. Therefore, the vast majority of stitching networks trained on synthetic datasets are not good enough for realistic stitching scenes.

To overcome the limitations of supervised learning-based image stitching, some unsupervised deep learning-based methods have been proposed [30]. The existing unsupervised deep learning-based models can get better stitching effects. However, most of these networks improve the accuracy of image stitching by multiple cost volume calculations or deepening the network depth, which is very time-consuming. As we know, in many practical application scenarios, the image stitching task needs to be completed in a short time, such as imaging stitching tasks for underwater robotics and unmanned aerial vehicles [31], [32]. This is one of the motivations of this study.

To deal with the problems in image stitching, a fast unsupervised image stitching network (FUISNet) is proposed in this article. Unlike traditional strategies that optimize model results by increasing network depth, our network utilizes an efficient feature extraction module (FEM). In addition, our network replaces the method of calculating the cost volume used in traditional stitching models with an efficient homography estimation module (HEM), which significantly reduces memory usage while improving stitching performance.

The main contributions of this article can be summarized as follows.

1) An improved unsupervised learning-based image alignment network for image stitching tasks is designed, where a deformation adaptive FEM and an efficient HEM are proposed to improve the speed, accuracy, and robustness of the image stitching model.
2) A novel loss function for the image stitching task is presented, where not only the overall differences between pixels are taken into account, but also the optimization of edge content is specifically enhanced.
3) A stitching restoration network with less number of parameters is proposed to achieve better stitching results.

This article is organized as follows. Section II gives an overview of the work related to the field of image stitching. In Section III, the proposed method and its design principles are presented. In Section IV, the experimental procedure is presented and the proposed model is compared with classical and state-of-the-art image stitching methods. The ablation study for the improvement of the proposed model and some discussions are given in Section V. Finally, a conclusion is given in Section VI.

II. Related Works

In this section, we introduce the work related to three mainstream image stitching methods, including image stitching methods based on feature point matching, image stitching methods driven by seam lines, and image stitching methods based on deep learning.

A. Based on Feature Point Matching

In the early stage, most of the image stitching algorithms are based on the direct alignment of images by pixel intensity. So far, the most mature image stitching approach is the method based on the underlying feature point matching. In 2006, Brown and Lowe [33] presented a pioneering work on this stitching theory, using automatic stitching methods with SIFT [34], random sample consensus (RANSAC) [35] and multiband mixing. However, for images with parallax, a single homography transformation model is often difficult to achieve accurate alignment.

To solve this problem, researchers have proposed some new ideas for aligning images by combining multiple parametric alignment models. For example, Zaragoza et al. [36] proposed an as-projective-as-possible (APAP) method that used a movable direct linear transformation (DLT) model to estimate the transformation parameters and introduced a framework similar to energy optimization to achieve the most natural stitching effect possible. This method is able to adaptively stitch images onto a common plane and can handle different image distortions during the stitching process. However, the APAP algorithm has some problems when dealing with large parallax images because it tries to minimize the planar perspective distortion on the whole image. Thus, some improved models have been proposed [37], [38].

Although traditional algorithms based on underlying feature point matching have been used in image stitching for many applications, these methods inherently rely on low-level feature detection methods. So these models may not adequately describe the complex content and structure of an image, leading to a high risk of false matches when performing feature matching. In addition, these low-level feature detection methods may be sensitive to changes in the illumination, scale, and perspective of the image.

B. Driven by Seam Lines

The seam line-driven image stitching method is also a common stitching technology, which obtains a natural stitched image by finding seam lines [39]. For example, He et al. [40] proposed a seam-driven image stitching method, which is a region-based seam optimization method by trying to find the optimal seam. Liao et al. [41] proposed a method based on iterative seam estimation, in which the initial seam is estimated by conventional seam cutting, and the pixels on the seam are evaluated using a hybrid quality assessment method.

Compared with methods based on feature point matching, seam-driven methods do not require feature point extraction and matching and thus may perform better for problems where some feature points are not obvious or difficult to match. However, if there is a significant change in viewpoint in the scene, resulting in a large change in the geometric relationship between the images, it may be difficult to align the images accurately. So that the seam-driven image stitching method

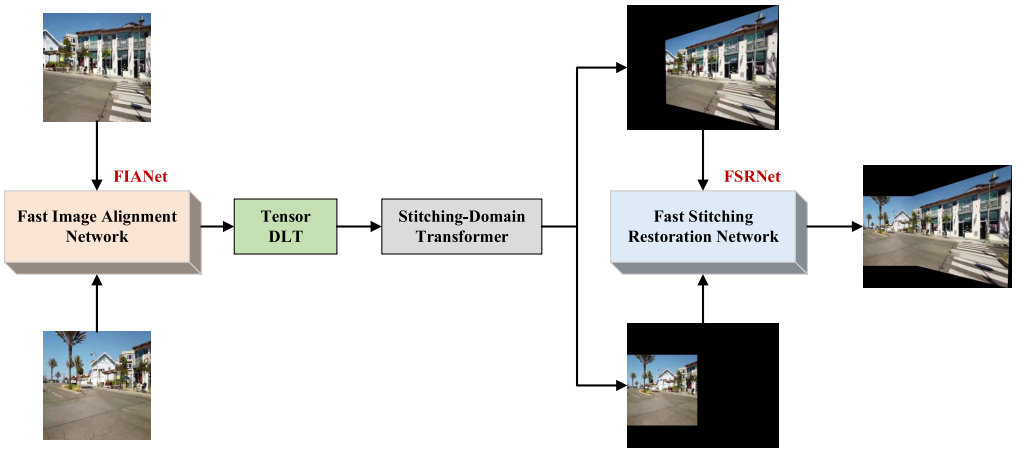


Fig. 1. Overall structure of the proposed FUISNet. There are two main parts: the FIANet and the FSRNet. DLT means the direct linear transformation operation.

may not be able to handle the situation, resulting in stitching failure.

C. Based on Deep Learning

In recent years, deep learning-based stitching methods have gradually become an important direction in the field of image stitching, because deep learning-based methods can learn higher-level and abstract features and better take into account the global information of images [42], [43], [44]. Homography estimation is closely related to image stitching, which is a technique used to compute transformations between images. In deep learning-based homography estimation [24], [25], [45], CNNs are usually used to extract features of images and these features are used to estimate the homography matrix between two images. For example, Hoang et al. [46] used feature points extracted by neural networks for alignment to compare patch similarity metrics in images to determine constraints between image pairs. Zhao et al. [47] proposed a deep neural network for stitching small parallax images, employing a coarse-to-fine homography estimation.

To more closely match the needs of real-world scenes, Nie et al. [30] proposed an unsupervised deep learning image stitching model consisting of coarse alignment and image reconstruction to achieve image stitching of arbitrary resolution and arbitrary viewpoint. For image stitching with fixed viewpoints, Song et al. [48] proposed an end-to-end image stitching network with a multihomography estimation scheme to address the parallax distortion due to depth differences in the scene. Kweon et al. [49] proposed a new deep image stitching framework that uses pixel-wise warp fields to handle large-parallax problems. Nie et al. [50] presented a parallax-tolerant image stitching model (PTISNet) that achieves image registration from global homography to local thin-plate spline motion through joint optimization of alignment and distortion.

As introduced above, researchers optimize their stitching results usually through the strategy of increasing the network depth. However, as the network depth increases, the computational complexity of the model increases significantly, thus increasing the training and inference time. Also, deeper networks require more memory to store more parameters and intermediate computational results. So, the focus of this article is how to optimize the structure of deep learning-based stitching networks while reducing the computational complexity without compromising the final stitching effect.

III. Proposed Image Stitching Method

In this article, an improved deep learning-based image stitching network model is proposed. The main purpose of this study is to solve the problem of poor timeliness of traditional deep learning-based stitching methods. The overall structure of the proposed FUISNet is shown in Fig. 1.

As depicted in Fig. 1, the proposed FUISNet comprises two main components: the fast image alignment network (FIANet) and the fast stitching restoration network (FSRNet). FIANet is primarily responsible for calculating the corresponding transformation relationship between the input image pairs based on the homography matrix. It accomplishes this by distorting the input image pairs through the stitching-domain transformer. The distorted image pairs are then stitched together by FSRNet. In addition, FSRNet addresses any resulting seams and artifacts, producing a stitched image with superior final results. In summary, compared to the existing models based on deep learning, our proposed model is a fast unsupervised image-stitching model that does not require labeled data for training and is specially designed to process image-stitching tasks quickly. The proposed FUISNet will be introduced in detail as follows.

A. Fast Image Alignment Network

FIANet mainly consists of two parts, the feature pyramid and the HEM, as shown in Fig. 2. Detailed depictions of the above proposed FIANet network are as follows.

1) Feature Extraction Module: As shown in Fig. 2, after the images are stacked and fed into our network, they are extracted by a modified YOLO-based backbone structure for feature extraction [51]. In the proposed FEM, a repeated network hierarchy assignment strategy is presented, to provide a more detailed decomposition of the inputs and stabilize the propagation of the gradient while maintaining the continuity of the features.

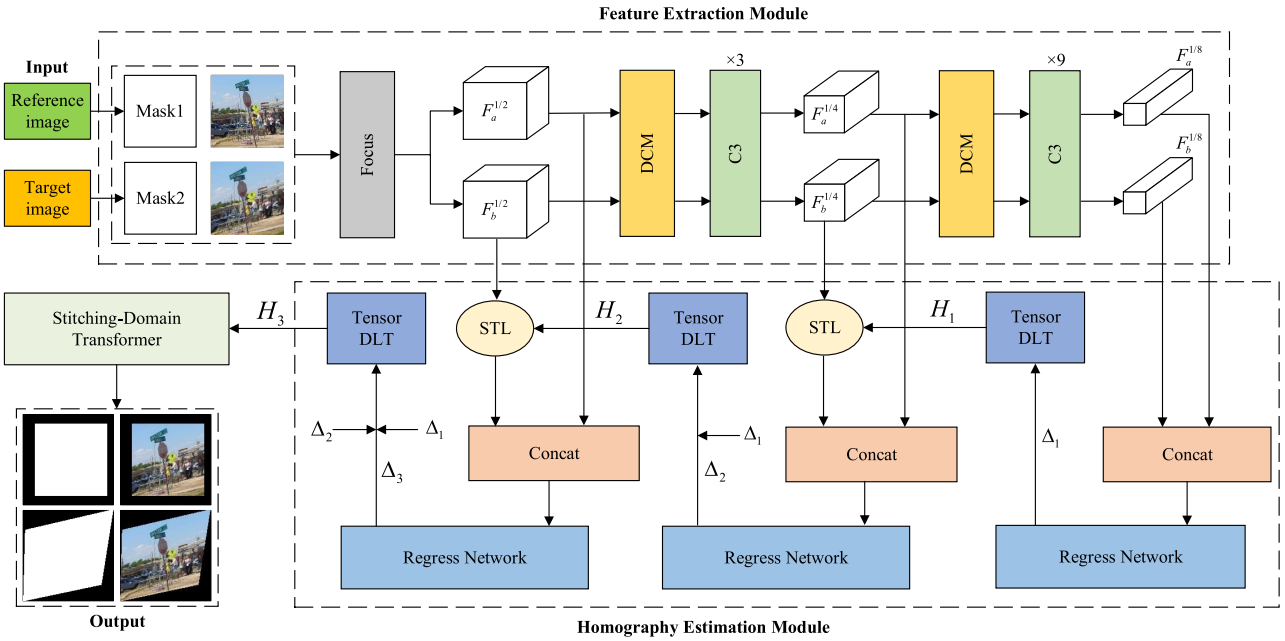


Fig. 2. Overall structure of the FIANet, where "STL" means warping using spatial transformer layer, "Concat" means merging feature maps with Concat operation, and "DCM" means deformable convolution module. Input1 (reference image), Input2 (target image), Mask1, and Mask2 are of the same size and are stacked as inputs to the whole network. The input is passed through the first part, i.e., the FEM based on an improved YOLO-based network with a feature pyramid. This module employs YOLO's focus and C3 modules to make the network have better feature representation capability and higher computational efficiency. In addition, an integrated DCM is introduced into the FEM module, to improve feature extraction capability for deformation targets. Then, the obtained feature maps at different scales are fed into the second part, i.e., the HEM for multiple aggregations, regression calculation, and twisting operations to achieve homography estimation from coarse to fine at the feature level.

Specifically, the feature maps of the two input images obtained at different scales, namely, $F_a^{1/2}$, $F_a^{1/4}$, $F_a^{1/8}$, $F_b^{1/2}$, $F_b^{1/4}$, and $F_b^{1/8}$ are used to construct the feature pyramid [52], [53]. Then, the homography matrix is estimated from high level to low level.

As we know, in practical applications, the input images may have some distortions due to various reasons, such as changes in camera pose, lighting conditions, and so on. To deal with this problem, in this article, we add an integrated deformable convolution module (DCM) [54], [55] into the FEM module to expand the perceptual field of the model and to improve the network's adaptability for deformations. The workflow of the DCM module is shown in Fig. 3.

As shown in Fig. 3, assuming that the original input feature map is x, and the convolution kernel is w, the position of the traditional convolution operation on the location of p can be expressed as

$$y(p) = \sum_{p_n} w(p_n) \cdot x(p + p_n) \tag{1}$$

where p is the output position of the convolution operation and p_n is the relative position in the convolution kernel w.

For the deformable convolution operation with resolution $k \times k$, the feature map is first passed through a traditional convolution operation to obtain a mask map of $k \times k \times 3$, and this mask map is divided into an offset vector and a weight mask as follows:

$$\text{Conv-op}(x) = \Delta P + \Delta M \tag{2}$$

where ΔP is the offset vector; ΔM is the weight mask; and Conv-op(x) is the traditional convolution operation.

For each position in the input feature map, we offset it using the offset vector ΔP to get a new position. The weight mask ΔM is constrained to a value between $[0, 1]$ by the Sigmoid function, which gives different weight values to the convolution points at each location, providing more freedom for sampling.

The overall output of the DCM is described as follows:

$$y(p) = \sum_{p_n} w(p_n) \cdot x(p + p_n + \Delta P) \cdot \text{Sigmoid}(\Delta M). \tag{3}$$

2) Homography Estimation Module: In general, during the alignment phase, most of the methods compute the correlation matrix between two feature maps to obtain the matching relationship between them [56]. However, these methods require a convolution and a large number of multiplication operations to calculate the correlation, which is very time-consuming and needs a lot of computer memory. To deal with this problem, a fast HEM based on an improved regress network is proposed in this article.

As shown in Fig. 2, we first use a simple Concat operation to merge the feature map pairs obtained from the feature pyramid in the channel dimension, which is described as follows:

$$F_c = F_a \oplus F_b \tag{4}$$

where $F_a \in w \times h \times c$ and $F_b \in w \times h \times c$ are the two input feature maps and $F_c \in w \times h \times 2c$ is the output after the Concat operation $\oplus$.

Then, an improved regression network is used to compute the correlations of F_c. The proposed regression network consists of three convolutional layers (Conv), one depth-wise

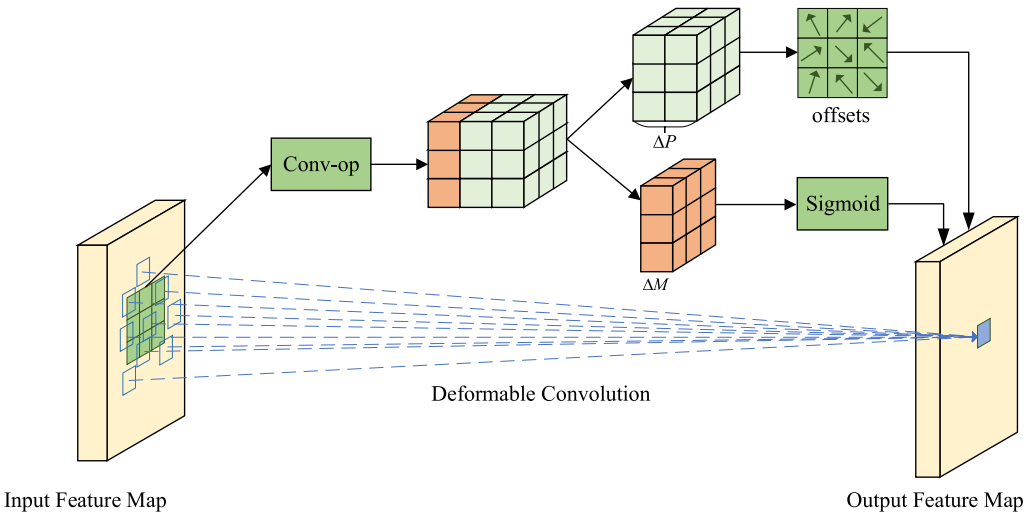


Fig. 3. Workflow of the DCM module, where $_1P$ is the offset vector; $_1M$ is the weight mask; Conv-op denotes the traditional convolution operation; and Sigmoid is the sigmoid function.

separable convolutional layer (DW-Conv), and two fully connected (FC) layers [57]. Among them, DW-Conv is used to downsample the feature map instead of the average pooling layer. The main reason for using DW-Conv is that it can retain more positional information by shifting the convolution kernel when downsampling and learning the relationship between different feature map channels. To make the model have better gradient propagation properties, the SiLU activation function is used in this article, which is described as follows:

$$f(x) = x \cdot \frac{1}{1 + e^{-x}}. \tag{5}$$

Given that the four corner coordinates of the image are (u_1, v_1), (u_2, v_2), (u_3, v_3), and (u_4, v_4), the offsets of these four points are denoted as $(\Delta u_1, \Delta v_1)$, $(\Delta u_2, \Delta v_2)$, $(\Delta u_3, \Delta v_3)$, and $(\Delta u_4, \Delta v_4)$. The coordinate offsets Δ_n are output by the regression network, namely,

$$\Delta_n = \begin{pmatrix} \Delta_{u_1} & \Delta_{v_1} \\ \Delta_{u_2} & \Delta_{v_2} \\ \Delta_{u_3} & \Delta_{v_3} \\ \Delta_{u_4} & \Delta_{v_4} \end{pmatrix}, \quad n \in \{1, 2, 3\}. \tag{6}$$

These offsets uniquely determine the corresponding homography matrices. After obtaining the feature maps F_b and the homography matrices generated by the upper layer (see Fig. 2), we warp the feature maps using spatial transformer layer (STL) [58]. In the STL layer, a uniform grid is generated for the input F_b first. Then, transform the grid using the homography matrices H_n. Finally, the Bilinear interpolation is used to extract the pixel values according to these new positions, and the transformed feature map F_b' with the same size of input feature map is generated and transferred to the lower layer. At last, the 3×3 homography matrice H_3 is obtained.

3) Objective Function: In the training of the depth homography estimation, most of the unsupervised learning-based methods use the loss function as follows [23], [24], [59]:

$$L_p = L_1\{I_a, \mathcal{W}(I_b)\} \tag{7}$$

where I_a and I_b denote the input reference image and the target image, respectively; $L_1\{\cdot\}$ denotes the calculation of the average absolute error loss; and $\mathcal{W}(\cdot)$ denotes the use of estimated homography to distort an image to align with another image.

Based on this type of loss function, in order to make the distorted target image close to the reference image, the reference image and the distorted target image are compared in whole. However, for a pair of images that need to be stitched together, they will not completely contain each other. So the alignment operation in the whole image does not make much sense for the final result, and it will extend the run time.

To solve this problem, an improved alignment loss for the overlapping regions is proposed in this article, which is defined as follows:

$$\begin{aligned} L_p' = {} & \lambda_1 \cdot L_1\{I_a \odot M_{\text{overlap}}, \mathcal{W}_1(I_b) \odot M_{\text{overlap}}\} \\ & + \lambda_2 \cdot L_1\{I_a \odot M_{\text{overlap}}, \mathcal{W}_2(I_b) \odot M_{\text{overlap}}\} \\ & + \lambda_3 \cdot L_1\{I_a \odot M_{\text{overlap}}, \mathcal{W}_3(I_b) \odot M_{\text{overlap}}\} \end{aligned} \tag{8}$$

where $\odot$ denotes the interpixel multiplication; λ_1, λ_2, and λ_3 are the weight factors of the loss items of different layers; and $\mathcal{W}_i$ denotes the distortion operation corresponding to the homography matrix obtained from the feature maps at different scales and $i \in \{1, 2, 3\}$. M_{overlap} is the calculated overlap area, which is described as follows:

$$M_{\text{overlap}} = M_a \& \mathcal{W}(M_b) \tag{9}$$

where & denotes the "And" operation; and M_a and M_b denote the all-ones mask with the same size as the reference image I_a and the target image I_a, respectively.

In addition, to consider the image content in the image stitching, a content loss function is defined in this article, namely,

$$L_c = L_1\{\text{Canny}(I_a'), \text{Canny}(\mathcal{W}_i(I_b'))\} \tag{10}$$

where $i = 1, 2, 3$; Canny$(\cdot)$ denotes the Canny operator. Since this loss function gives more weight to the image edges, it tends to align the edges first, which favors the stitching results.

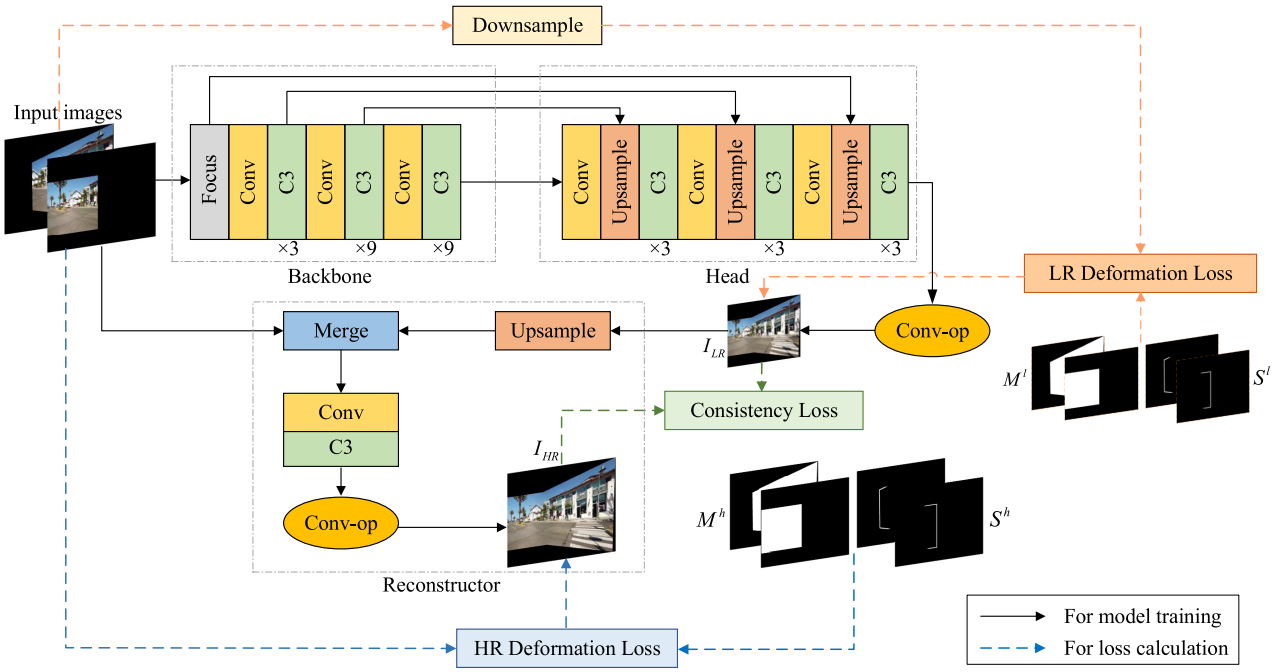


Fig. 4. Overall structure of the FSRNet, which consists of a backbone network, a head network, and a reconstructor. "Conv-op" means the convolution operation; I_{LR} and I_{HR} denote the stitched image of the low-resolution (LR) and high-resolution (HR), respectively. M^l and M^h are the image masks of LR and HR, respectively. S^l and S^h are the seam masks of LR and HR, respectively. The HR deformation loss is calculated by combining the input with M^h, S^h, and the image I_{HR}. In the LR branch, the input is resampled by downsampling with M^l, S^l and the image I_{LR}. After obtaining the HR and LR reconstructed images, the consistency loss of both is finally calculated.

The final loss for the homography estimation is formulated as a weighted sum of the above loss functions

$$L = \lambda_p L'_p + \lambda_c L_c \tag{11}$$

where λ_p and λ_c are the weights for the loss function L'_p and L_c, respectively. In this study, to balance pixel alignment and structural alignment during the training process, λ_p and λ_c are both set to 1.

Remark 1: Based on the proposed loss function, only the differences between pixels in the overlapping regions are used for alignment calculation to improve the working efficiency. In addition, to better capture the subtle structural differences, edge content is combined for alignment to ensure higher quality alignment results in critical edge and structural regions.

B. Fast Stitching Restoration Network

Given that a single homography cannot fully describe complex spatially transformed images with parallax depth variations, it is difficult to fully align the input images in the real scene dataset during the image alignment phase. In order to diminish the pixel-level misalignment caused by the alignment phase, many researchers have used generative networks to perform restoration reconstruction of prealigned images [60]. To improve the quality and accuracy of the reconstructed images, most researchers have tried to preserve as much detailed information as possible by deepening the network strategy. Although a natural stitched image is generated in the end, a large amount of computational resources are sacrificed.

To better solve the above problems, an efficient fast image stitching restoration network is introduced, as shown in Fig. 4, which is capable of processing stitched prealigned image pairs and restoring the resulting seams, artifacts, and other problems. Meanwhile, we satisfy the demand for real-time stitching applications by further reducing the number of network parameters. The detailed process of the proposed image stitching restoration network is described below.

1) Network Structure: The overall network structure of our proposed fast image stitching restoration network consists of a backbone network, a head network, and a reconstructor. The backbone and head networks are mainly responsible for the processing of LR branch images, while the reconstructor is mainly used to combine LR branches to reconstruct HR images.

The network takes the distorted image pairs of the same size generated in the alignment phase as input. In the backbone network, an improved focus module is proposed and used first, which extracts features by performing focus operations on the input. The structure of the proposed focus module is shown in Fig. 5. Compared with the conventional focus module, the improved focus module is able to handle the case of two inputs.

After the focus module, it is a Conv module (each Conv module consists of a convolutional layer, a batch normalization (BN) layer, and a ReLU activation function layer). This Conv module performs a convolution operation and downsamples the output to 1/2 the size of the original image to obtain the feature map $P_a^{1/2}$.

Subsequently, after the combination of another Conv module and C3 modules, the feature map $P_a^{1/4}$ is obtained. Then, a series of convolution and stacking of C3 modules are performed to obtain the feature maps $P_a^{1/8}$ and $P_a^{1/16}$. The number of channels of the feature maps is 64, 128, 256, and

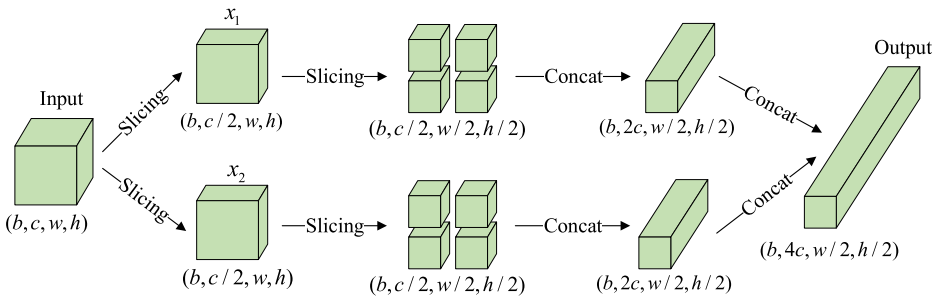


Fig. 5. Structure of the proposed focus module, where the input dimension is ▷b,c,w,h◁ and the shape of the output tensor is ▷b, 4c, $w/2$, $h/2$◁.

512 in order, and these feature maps play an important role in the subsequent processing of the network.

In the head network, the output of the previous stage is first passed through a Conv module with a convolutional kernel of size 1×1 to reduce the number of feature channels to 256. The feature maps $P_b^{1/8}$ are obtained by upsampling operation using nearest neighbor interpolation and then concatenating with feature maps $P_a^{1/8}$ of the same size as the backbone network.

Similarly, the feature maps of $P_a^{1/4}$ and $P_a^{1/2}$ are obtained sequentially through a series of combinations of Conv and C3 modules, and a series of concatenating operations are performed. Finally, the Reconstructor is used to aggregate the LR images and perform an HR image reconstruction.

2) Objective Function: Like the alignment network, the FSRNet is also trained in an unsupervised manner. During the training process, we use image masks and seam masks to learn the deformation rules of image stitching (see Fig. 4).

By using the image masks, the features of the reconstructed image can be constrained to be close to the prealigned image. By introducing explicit spatial constraints, we can ensure that the fusion result remains highly consistent with the input image in terms of structure and texture. In addition, these masks explicitly define the weight of each input image in the fusion result, allowing us to maintain or emphasize the features of a particular input image within a specific region.

On the other hand, seam masks are used to constrain the edges of the overlapping regions to produce a natural effect. This constraint ensures that the fused content within the seam region transitions smoothly with its surroundings, eliminating discontinuities or visual artifacts. In particular, the seam mask focuses on the fused region, ensuring that the content within that region is as close as possible to the original input image.

Specifically, the fusion-aware loss function is used in this article [30]. In the process of calculating the loss function, the first step is to process through the image mask to obtain the corresponding seam mask. For the input mask M, its seam mask can be obtained as follows.

First, the image edge mask is calculated by

$$M_{\text{edge}} = \mathcal{C}\{\nabla M_x + \nabla M_y\} \tag{12}$$

where $\mathcal{C}\{\cdot\}$ is the operation of clipping all elements to between 0 and 1; and ∇M_x and ∇M_y are the two gradient values, which are obtained by

$$\nabla M_x = \mathcal{C}\{|M * G_x|\}, \nabla M_y = \mathcal{C}\{|M * G_y|\} \tag{13}$$

where $*$ stands for the convolution operation; M denotes the input image mask; and G_x and G_y are the Sobel operators in the X- and Y-direction, respectively [61].

Then, the seam mask M_{seam} can be obtained by performing three expansion operations on M_{edge}, as follows:

$$M_{\text{seam}} = M_{\text{edge}} * E_{3\times3} * E_{3\times3} * E_{3\times3} \tag{14}$$

where $E_{3\times3}$ is an all-ones matrix of 3×3.

The fused perceptual loss function is divided into three main parts: HR deformation loss ($\mathcal{L}_{\text{HR}}$), LR deformation loss ($\mathcal{L}_{\text{LR}}$), and consistency loss ($\mathcal{L}_{\text{CS}}$).

The loss function of the LR deformation branch can be expressed as

$$\mathcal{L}_{\text{LR}} = \lambda_i \mathcal{L}^l_{\text{image}} + \lambda_s \mathcal{L}^l_{\text{seam}} \tag{15}$$

where $\mathcal{L}^l_{\text{image}}$ and $\mathcal{L}^l_{\text{seam}}$ denote the resulting image loss and seam loss at LR, respectively; and λ_i and λ_s represent the weights of each component.

In the same way, the loss function of the HR deformation branch can be expressed

$$\mathcal{L}_{\text{HR}} = \lambda_i \mathcal{L}^h_{\text{image}} + \lambda_s \mathcal{L}^h_{\text{seam}} \tag{16}$$

where $\mathcal{L}^h_{\text{image}}$ and $\mathcal{L}^h_{\text{seam}}$ denote the resulting image loss and seam loss at HR, respectively.

In this study, to pay more emphasis on eliminating visible seams in the merged or stitched portion, λ_i and λ_s are set to 1 and 10 in the training process of FSRNet, respectively. The main reason for this setting is that appropriately increasing the weight of seam correction can help achieve a more natural color transition when repairing and reconstructing stitched images [30].

In addition, in order to better fuse the LR and HR ground reconstruction results, content consistency loss is used, namely,

$$\mathcal{L}_{\text{CS}} = \mathcal{L}_1(I_{\text{HR}}/\sigma, I_{\text{LR}}) \tag{17}$$

where I_{HR} and I_{LR} denote the HR and the LR images, respectively; $\mathcal{L}_1$ is the L1 loss function; and σ is defined as follows:

$$\sigma = \frac{W}{w} \tag{18}$$

where W and w are the widths of the stitched HR and LR images, respectively.

In summary, the objective function of FSRNet is defined as follows:

$$\mathcal{L}_R = \mathcal{L}_{\mathrm{CS}} + \mathcal{L}_{\mathrm{LR}} + \mathcal{L}_{\mathrm{HR}}. \tag{19}$$

Finally, the results are obtained by multiplying the loss of each component by the batch size and returning the total loss and the value of each loss as the result.

The workflow of the whole deep network-based image stitching proposed in this study is as follows.

Step 1: The image pairs to be stitched are input to the FIANet proposed in this study.

Step 2: Using the improved YOLO-based feature pyramid, the features of the image pairs are extracted, and feature maps at three different scales are obtained.

Step 3: Obtain the homography matrix representing the mapping relationships between images by the HEM.

Step 4: Use the computed homography matrix to distort the input image pairs by stitching-domain transformer module to generate coarse-aligned distorted image pairs.

Step 5: The distorted image pairs are input to the FSRNet proposed in the text for the final stitching reconstruction to generate stitched images with natural transitions.

IV. Experiment

A. Dataset and Evaluation Criteria

In our experiments, we used two datasets. The first is the Warped MS-COCO dataset [62], which is a public dataset commonly used for homography estimation tasks. The process of generating the synthetic dataset is: first, cropping a portion from the COCO dataset as the reference image, with its four corners defining the initial bounding box. This bounding box is then translated and randomly perturbed to ensure it remains a convex quadrilateral. Next, the homography matrix is calculated to apply a perspective warp to the original image, creating the transformed target image. Pixels outside the second frame are set to black to generate the ground truth image. In experiments, we synthesize a total of 50 000 image pairs as the training set and 5000 image pairs as the test set. In addition, to consider the higher stitching resolution in the synthetic dataset, we also generate a dataset of 1000 pairs with a resolution of 256 × 256 containing real stitching results to test the effect of various stitching solutions on the synthetic dataset. For the specific task of image stitching, the warped MS-COCO dataset is capable of modeling the imperfect alignment and minor distortions that may exist between images in the real world.

The second is the UDIS-D dataset [30], which is obtained from variable motion videos. By extracting the frame images with different intervals from these videos, the image pairs with different overlap rates and different parallax degrees are obtained. This real-world scenario dataset includes a variety of scenarios, including indoor, outdoor, dark environment, low texture, small foreground, and large foreground. The dataset contains 10 440 pairs of training images, and 1106 pairs of test images. As a real scene dataset, the UDIS-D dataset is closer to the actual application scene, which can reflect the challenges in various real-world scenarios, such as images captured from different viewpoints and different lighting conditions.

In terms of evaluation metrics, peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and RMSE are used as the main evaluation criteria, where PSNR and SSIM are the most commonly used evaluation metrics in assessing image quality [63], while RMSE is commonly used in homography estimation tasks to predict the error between the homography transform and the true homography transform [59].

RMSE can be calculated by

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\Delta_i^{\mathrm{pred}} - \Delta_i^{\mathrm{gt}}\right)} \tag{20}$$

where Δ_i^{pred} and Δ_i^{gt} denote the offsets of the predicted and true four coordinates in the X- and Y-direction, where $N = 8$ in this study. The smaller the RMSE value is, the closer the estimation result is to the true value.

PSNR is calculated as follows:

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{\mathrm{MAX}^2}{\mathrm{mse}}\right) \tag{21}$$

where MAX is the maximum possible pixel value of the image, which is set to 255 in this article; mse is calculated by

$$\mathrm{mse} = \frac{1}{mn}\sum_{i=1}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2. \tag{22}$$

SSIM is defined as follows:

$$\mathrm{SSIM}(x,y) = \left[l(x,y)^{\alpha} \cdot c(x,y)^{\beta} \cdot s(x,y)^{\gamma}\right] \tag{23}$$

where $l(x,y)^{\alpha}$, $c(x,y)^{\beta}$, and $s(x,y)^{\gamma}$ are calculated based on the correlation between samples X and Y with respect to luminance, contrast, and structure, respectively. In this study, we take $\alpha = 1$, $\beta = 1$, and $\gamma = 1$.

Remark 2: In the comparison of homography estimation on the synthetic dataset MS-COCO, RMSE is used because the synthetic dataset includes the true homography matrix for each image pair. However, in the real-world UDIS-D dataset, the true homography matrix is not available. Therefore, for real-world datasets, PSNR and SSIM of the overlapping regions between stitched images are used to evaluate the alignment accuracy. In comparison to stitching results, the stitched images generated by all methods are first resized consistently. Then, the two metrics PSNR and SSIM are calculated sequentially for the entire test set and finally averaged, to evaluate the stitching task.

B. Experimental Details

In this study, FUISNet is implemented using the PyTorch framework, with training and testing performed on a single NVIDIA GTX 2080. The batch size for training is 4. The FIANet is trained with 150 epochs using the synthetic dataset and fine-tuned to 50 epochs on the UDIS-D dataset using Adam. The initial learning rate is set to 0.0001 and the momentum factor is set to 0.9. Then, when training FSRNet, 30 epochs are trained with UDIS-D, the initial learning rate is set to 0.0003, and the momentum factor is set to 0.9.

TABLE I
RMSE (↓) OF THE HOMOGRAPHY ESTIMATION ON THE SYNTHETIC DATASET MS-COCO

Method	SR [64]	DHN [59]	UDHN [23]	CUDHN [24]	UDISNet [30]	FIANet (Ours)
0~30%	0.71	3.38	2.33	15.23	1.21	1.77
30%~60%	1.26	4.91	3.75	18.37	1.50	1.78
60%~100%	19.62	7.88	6.51	21.21	3.10	1.82
All	8.45	5.64	4.43	18.57	2.06	**1.79**

TABLE II
PSNR (↑) AND SSIM (↑) OF THE OVERLAPPING REGIONS ON THE REAL DATASET UDIS-D

Method	PSNR				SSIM			
	0~30%	30% ~ 60%	60%~100%	All	0~30%	30%~60%	60%~100%	All
SR [64]	24.35	21.48	17.24	20.64	0.79	0.70	0.50	0.65
DHN [59]	15.46	12.85	11.16	12.95	0.42	0.15	0.07	0.19
UDHN [23]	18.86	16.03	13.21	15.75	0.55	0.3	0.19	0.33
CUDHN [24]	16.37	13.86	11.30	13.59	0.39	0.15	0.05	0.18
UDISNet [30]	26.25	23.37	19.93	22.85	0.89	0.79	0.59	0.74
PTISNet [50]	30.19	25.84	21.57	25.43	0.93	0.87	0.73	0.83
FIANet (Ours)	26.91	26.73	25.85	**26.54**	0.87	0.84	0.82	**0.86**

The training is conducted in an unsupervised way, which means that FUISNet only needs reference and target images as input without any labels. In addition, in order to make the model more generalizable, we perform data enhancement on the dataset before training FIANet, including the adjustment of saturation, hue, and brightness, as well as the operation of image rotation, translation, scaling, flipping, and perspective transformation.

In the comparison experiments in this study, to observe the performance of different stitching models on different overlap rates, the test sets are divided into three parts: 0%∼30%, 30%∼60%, and 60%∼100%, representing image overlap rate from high to low, like most methods (such as [27], [30]). The higher overlap makes image stitching easier.

C. Comparison of Homography Estimation

Due to the fact that image alignment is a crucial part of image stitching. To objectively evaluate the performance of our proposed alignment network, namely, FIANet, we compare it with the traditional stitching method SIFT + RANSAC (denoted by SR) [64], the supervised stitching method DHN [59], unsupervised stitching methods UDHN [23], CUDHN [24], PTISNet [50], and the alignment network used in UDISNet (called as UDISNet) [30]. SR is chosen as a representative traditional method because it outperforms most of the traditional solutions. DHN, UDHN, CUDHN, PTISNet, and UDISNet are deep learning-based solutions, where DHN is a supervised learning-based solution, and UDHN, CUDHN, PTISNet, and UDISNet are unsupervised learning-based solutions. The main reason for selecting these models is that they are state-of-the-art methods for image stitching and the results on the same public datasets of these models are available.

1) On the Synthetic Dataset MS-COCO: The first comparative experiment on homography estimation is performed on the Warped MS-COCO dataset. The results are shown in Table I.

From Table I, we can see that FIANet (Ours) outperforms all of the above methods. Although it is not the best performer in the first 60% part of the test dataset compared with the traditional method, the results of our model in each part fluctuate very little and the change of the proposed method is about 2.7%, which is lower than that of the traditional method about 97.2% (relative value). In addition, the RMSE of the proposed model reduces by 13.1% (relative value), compared to the UDISNet model (the second-best model). These results fully demonstrate the high accuracy as well as the generalizability of our proposed alignment network.

2) On the Dataset UDIS-D: The second comparison experiment on homography comparisons is performed on the UDIS-D dataset, which contains a variety of different real scenarios. Since UDIS-D does not contain alignment result truth labels, the PSNR and SSIM of the overlapping regions are used to evaluate their alignment performance. The comparison results are shown in Table II.

The results in Table II show that, in the evaluation results of PSNR, our method improves by 16.1% compared to UDISNet (the third-best model) and 4.4% compared to PTISNet (the second-best model), respectively. In the evaluation results of SSIM, our method improves by approximately 16.2% compared to UDISNet and 3.6% compared to PTISNet, respectively.

Furthermore, our network performs equally well in all parts of the test dataset, with the worst part of the PSNR only decreasing by 3.9% compared to the best part. In contrast, the performance of the traditional method is extremely uneven

TABLE III
COMPARISON OF STITCHING RESULTS ON PSNR (↑), SSIM (↑), PARA (↓), AND TIME (↓) ON THE DATASET MS-COCO

Method	PSNR				SSIM				PARA (M)	TIME (s)
	0~30%	30%~60%	60%~100%	All	0~30%	30%~60%	60%~100%	All		
APAP [36]	29.17	29.05	29.01	29.07	0.30	0.28	0.28	0.29	-	2.181
APIS [33]	29.17	29.10	29.07	29.11	0.22	0.21	0.21	0.21	-	0.161
LUIS [8]	29.79	29.36	29.28	29.48	0.34	0.33	0.30	0.32	-	0.793
VFISNet [26]	30.28	30.27	30.19	30.27	0.45	0.44	0.43	0.44	62.3	4.622
EPISNet [27]	30.54	30.36	30.36	30.43	0.46	0.45	0.45	0.45	183.7	5.753
PTISNet [50]	30.62	30.57	30.54	30.57	0.46	0.45	0.45	0.45	91.3	4.969
UDISNet [30]	30.91	30.75	30.64	30.78	0.48	0.47	0.46	0.47	188.0	6.284
Ours	30.98	30.94	30.82	**30.91**	0.49	0.49	0.48	**0.48**	**23.4**	**1.728**

across all parts of the dataset, with a difference of about 29.2% between the worst and best parts.

In summary, the experimental results from these two datasets clearly demonstrate the superiority of our fast registration network in achieving image alignment.

D. Comparison of Stitching Results

To verify the superiority of our proposed FUISNet in image stitching, the stitching results of FUISNet are compared with some classic and state-of-the-art methods, including APIS [33], APAP [36], VFISNet [26], EPISNet [27], UDISNet [30], LUIS [8], and PTISNet [50]. Among them, APIS is an important enhancement of the SIFT + RANSAC method for stitching multiple panoramic images; APAP uses a moving DLT to maintain the accuracy and continuity of perspective projection to the maximum extent, which is also a widely used traditional method; VFISNet and EPISNet are both for arbitrary viewpoint image stitching methods based on supervised learning methods, but VFISNet is only applicable to stitching image pairs with LR. So we compress the image pairs input to VFISNet to obtain the output results. UDISNet is currently the most widely used depth method for the field of arbitrary viewpoint image stitching. LUIS is an image stitching method based on larynx ultrasound superpixel features, primarily applied to medical images without using deep learning. PTISNet, on the other hand, is a newly proposed image stitching method that leverages deep learning techniques.

1) On the Synthetic Dataset MS-COCO: First, we will test the different stitching methods using the synthetic dataset MS-COCO. Since each method produces stitching images of different sizes, we need to crop them according to their smallest outer rectangle before comparing them with the true values of the synthetic dataset. Then, they are resized to match the size of the images corresponding to the true values. To test the stitching performance of different models, the PSNR and SSIM on each part of the dataset MS-COCO are given out. In addition, the number of model parameters (denoted by PARA) and the inference time (denoted by TIME) of different models are compared, to show the efficiency of each model. The results are shown in Table III and Fig. 6.

As can be seen from Table III, our method significantly outperforms other stitching methods. Our method improves the PSNR by 6.1% (to APIS), 6.3% (to APAP), 2.1% (to VFISNet), 1.6% (to EPISNet), 0.4% (to UDISNet), 4.6% (to LUIS), and 1.1% (to PTISNet), respectively. Regarding the results of SSIM, our method improves by 2.1% compared to UDISNet (the second-best model). It shows the superiority of our network for the stitching task.

Since APIS, APAP, and LUIS are based on traditional stitching methods, their model parameters are not discussed here. As can be seen in Table III, our model has the lowest number of parameters among the deep learning-based image stitching methods, which is approximately 62.4%, 87.3%, 87.6%, and 74.3% less than VFISNet, EPISNet, UDISNet, and PTISNet, respectively. Although, the inference time used by our method is not better than the traditional method APIS, and LUIS, it is the shortest among all other stitching methods based on deep learning, with reductions of about 62.6%, 70.0%, 72.5%, and 65.2% compared to VFISNet, EPISNet, UDISNet, and PTISNet, respectively. Our model even outperforms the traditional method APAP, with a reduction of about 20.8% in inference time. The main reason is that our design of efficient FEMs can significantly reduce computational load. The results show that our model maintains high accuracy while substantially reducing model size and inference time.

In Fig. 6, it is obvious to see that our method produces results that are closest to the ground truth. In contrast, the VFISNet produces a significant blurring effect when stitching. Although both EPISNet and UDISNet are also high-quality stitching methods, they have a slightly higher probability of error in the alignment process compared to our method. In addition, since PTISNet does not rely on homography estimation for alignment, it cannot handle stitching issues involving perspective transformations in synthetic datasets very well. It can also be seen from the figure that both APIS and APAP do not work well in terms of accuracy with prealignment. Because LUIS is based on SIFT for prealignment processing, it is only slightly better than the APAP method and far less than all deep learning-based methods when dealing with data with random viewpoints and diverse categories.

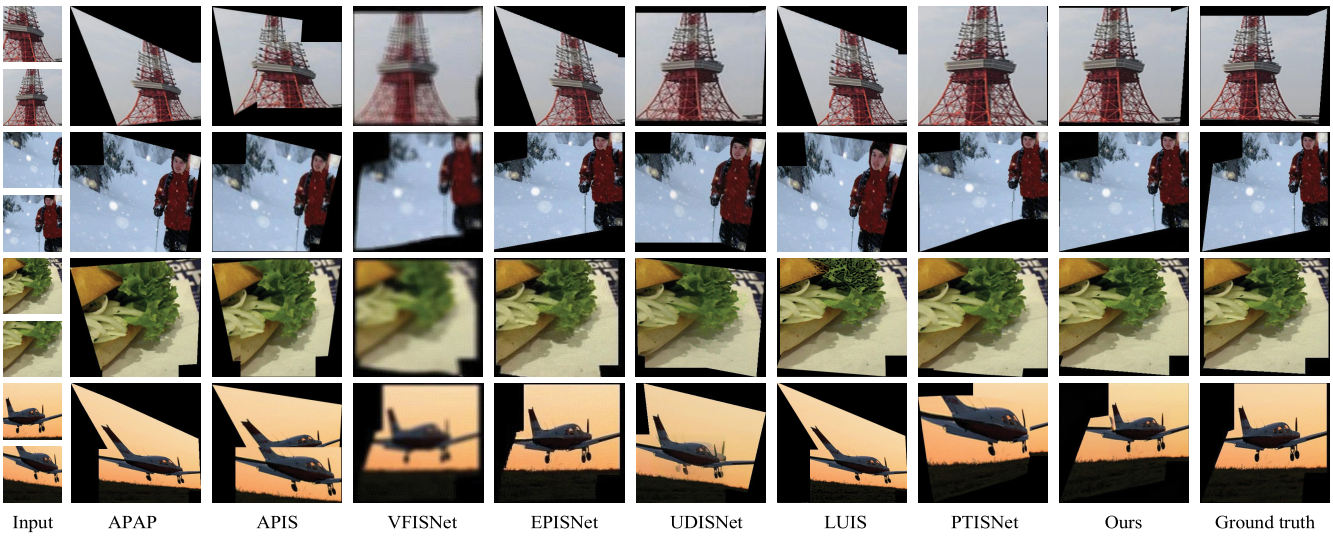


Fig. 6. Some examples of the stitching results on the dataset MS-COCO based on different models.

Remark 3: Although the improvements in PSNR and SSIM of the proposed model are not very dramatic, this is due to the fact that PSNR and SSIM are calculated on the whole image of the stitching task. Overall, our method has significant advantages in terms of stitching effectiveness and accuracy.

2) On the Dataset UDIS-D: In order to verify the performance of our network in real scenes, we also observe the output details of the above various stitching methods on the UDIS datasets. Because VFISNet can only handle small-resolution input image pairs, we perform twice bilinear interpolation operations on the output results of VFISNet. In this experiment, we observe the stitching results of various scenes, such as indoor, outdoor, dark environment, low texture, small foreground, and large foreground. Because the dataset UDIS-D is a real dataset, in this comparison experiment, we only give out the visual stitching results and focus on comparative analysis from the articulated part and overlapping part of the output images, respectively. The results are shown in Fig. 7.

From the comparison of the stitching results of different real scenes in Fig. 7, we can see that the APIS, APAP, LUIS, and EPISNet methods will have frequent color breaks in the articulated regions of the two input images, such as the stairs and walls in Fig. 7(a), the sky in Fig. 7(c), and the steps and floor tiles in Fig. 7(d). This phenomenon rarely occurs in VFISNet, UDISNet, and our proposed model. The main reason is that all these methods add reconstruction networks, which can further enhance the fusion effect between image pairs. Although APIS introduces an auto-correction strategy and LUIS uses a seamless connection strategy, they do not achieve a global natural color transformation. The EPISNet network has a reconstruction network, but it focuses on the preservation of edge structures within individual images and neglects the articulation part between images. Both PTISNet and our network can achieve smooth transitions in the articulated regions of the two input images.

For the overlapping part between images, APIS, APAP, and LUIS methods show serious ghosts, such as the stone in Fig. 7(b), and the pillar in Fig. 7(d), which is mainly due to the feature matching results of APIS, APAP, and LUIS that rely heavily on hand-made feature points. The introduction of the multiconstraint energy function strategy in LUIS also cannot get rid of the shortcomings of the traditional methods. The VFISNet method has a fatal problem in that it can only process the LR images. In addition, although the UDISNet network has better results in dealing with the articulated and overlapping regions of the images, the results of UDISNet are still blurred when it deals with textures in small foregrounds and dark scenes, such as the heater in Fig. 7(a), the garbage bin in Fig. 7(b), and the building in Fig. 7(c). This is mainly because in the UDISNet network, the input warped image pair is compressed and then combined with HR images together. During the compression process, a large amount of useful information is lost. PTISNet uses mask learning for reconstruction and stitching. Like our reconstruction network, it preserves the details of the original images during the reconstruction process.

Because the dataset UDIS-D is a real-world scenario dataset, to show the performance of the proposed model, we conducted an investigation of the subjective quality of the final image stitching results. In this investigation, the proposed FUISNet is compared with three other recent and well-performing methods, namely, UDISNet, LUIS, and EPISNet. The images produced by these methods are displayed as a set of images on the screen simultaneously in an anonymous and randomized manner. The users are asked to answer which of the set of images has the best stitching effects or whether the stitching of the two images is good or bad. Through our organized voting, a total of 58 people participated in the voting for the comparison with LUIS, 67 people for the comparison with UDISNet, and 69 people for the comparison with EPISNet. The proportion of those people with and without computer science background is about 1:1. The investigation results are shown in Fig. 8, which shows that the majority of users preferred the results produced by our proposed FUISNet.

In summary, our proposed network has the best stitching effect in handling various scenes. In addition, the results show that our network is robust and generalized, and it not only

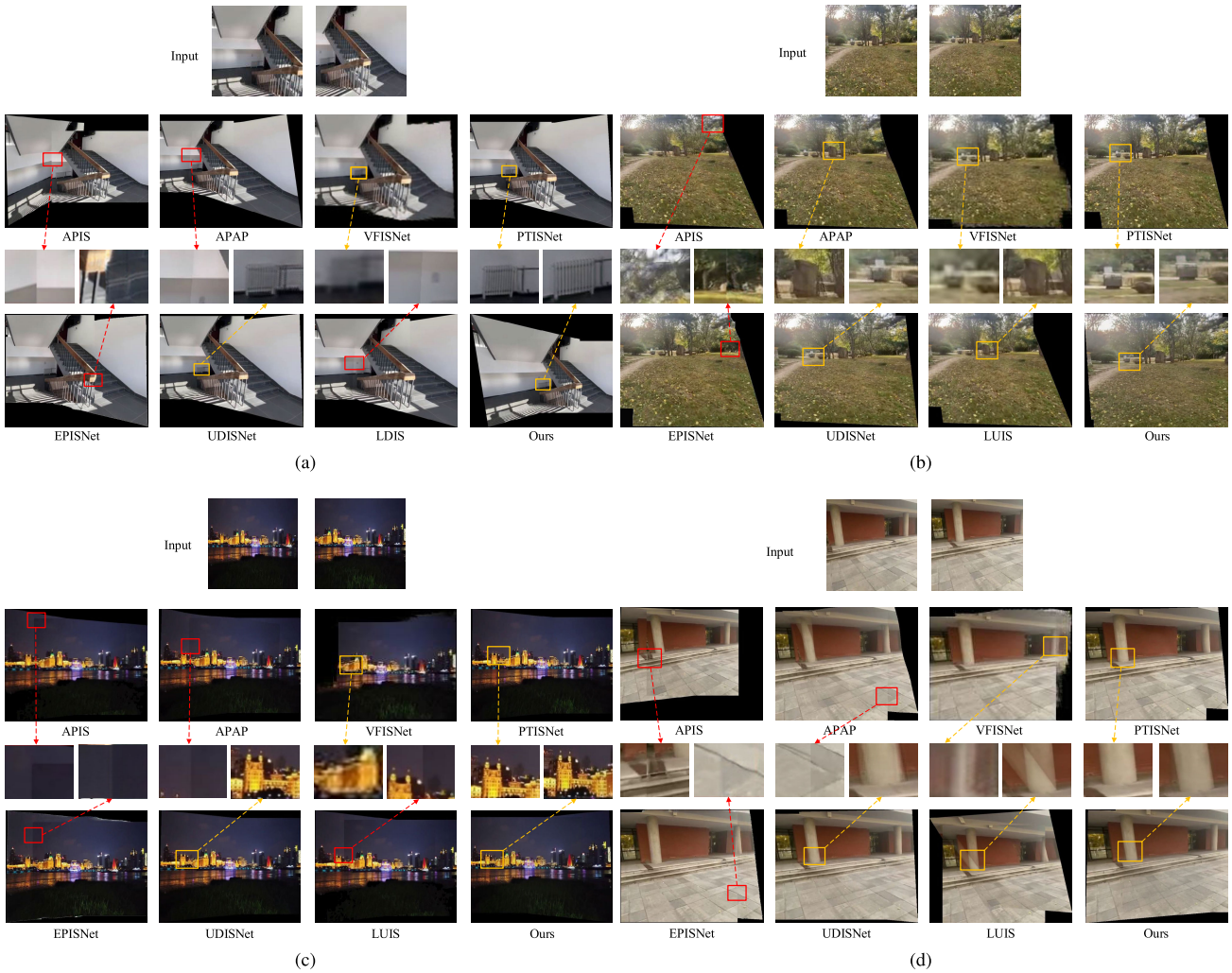


Fig. 7. Stitching effect comparison on the UDIS-D dataset. (a) On indoor scene. (b) On outdoor scene. (c) On dark scene. (d) On large foreground scene.

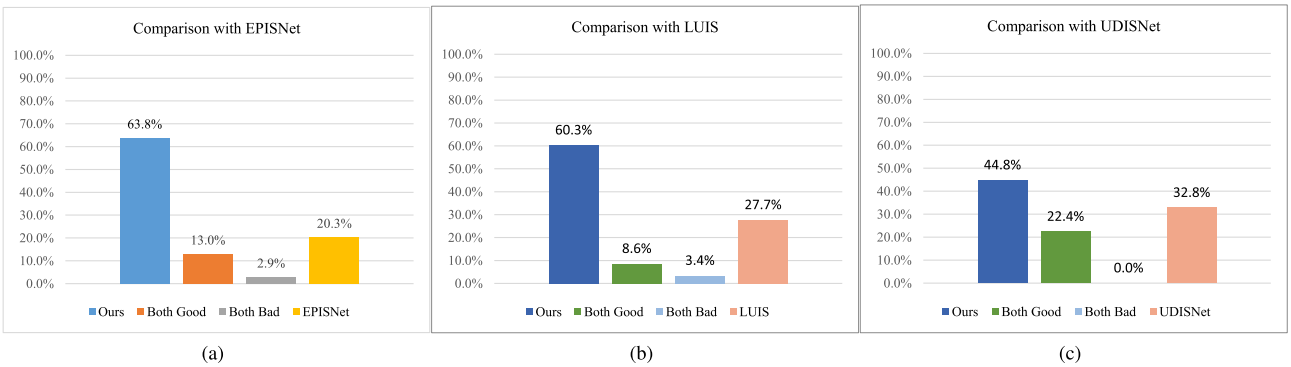


Fig. 8. Subjective evaluations of the stitching results based on the proposed model and other state-of-the-art methods. (a) Comparison with EPISNet. (b) Comparison with LUIS. (c) Comparison with UDISNet.

has an alignment network with high accuracy but also can reconstruct the image with sufficient detailed information to ensure better visual effects.

V. Discussion

The performance of the proposed model on the two public datasets has been proved by some comparison experiments in Section IV, especially the image alignment of the proposed model is tested and compared with other state-of-the-art methods. In this section, there is the first detailed analysis of the important hyperparameter selection of the alignment loss function. Then, the important improvement parts of the proposed model are discussed, where four decimal points are retained for RMSE. At last, the real-world application performance of the proposed model is discussed by a new real-world dataset.

A. About the Weights of the Alignment Loss Function

In the proposed FIANet, three weights (λ_1, λ_2, and λ_3) are introduced to represent the contribution of each layer to the

TABLE IV
RMSE (↓) OF THE HOMOGRAPHY ESTIMATION ON THE SYNTHETIC DATASET MS-COCO WITH DIFFERENT WEIGHTS OF THE ALIGNMENT LOSS FUNCTION

λ_1	λ_2	λ_3	RMSE
0	0	1	7.8879
0	1	1	2.3191
1	1	1	1.9053
16	4	1	1.7909
⋮	⋮	⋮	⋮
100	10	1	1.9810

TABLE V
RESULTS OF THE ABLATION EXPERIMENT ABOUT THE ALIGNMENT NETWORK

Models	Edge Loss	FEM	HEM	RMSE	TIME(s)	PARA(M)
UDISNet	×	×	×	2.0639	6.284	188.0
Model1	×	×	✓	2.2571	1.653	22.2
Model2	×	✓	✓	1.8034	1.728	23.4
Ours	✓	✓	✓	1.7909	1.728	23.4

final result [see (8)]. To determine the optimal values for these weights, some expanded homography estimation experiments on the synthetic dataset MS-COCO are conducted, where the settings of the proposed model are the same as the experiment in Section IV-C1, except choosing different values of weights. Since not all possible combinations can be tried, some most representative combinations are set up in this article. The results are shown in Table IV.

The results in Table IV show that the deeper the network is, the higher the accuracy of the prealignment is. In addition, since the highest layer of the network has the largest receptive field, the extracted features are the most global in nature. From Table IV, it can be seen that the best performance results are obtained when there is a correspondence between the hyperparameter combination selection and the feature map, namely, $\lambda_1 = 16$, $\lambda_2 = 4$, and $\lambda_3 = 1$.

B. About the Alignment Network

The improvements of the alignment network in the proposed model include FEM and HEM of FIANet, and the proposed edge loss function. In order to evaluate these improvements of the alignment network in this study, an ablation experiment on the synthetic MS-COCO dataset is conducted. In this ablation experiment, to facilitate the comparison of the accuracy of these alignment networks, the reconstruction network of all these models except UDISNet is replaced by the proposed FSRNet. The results of this ablation experiment for these models can be seen in Table V.

The results of Table V show that after using HEM, the number of parameters of Model1, Model2, and our model are reduced by 88.2%, 87.6%, and 87.6%, respectively, and the inference time is accelerated by 73.6%, 72.5%, and 72.5%, compared to UDISNet. It indicates that the use of the concat operation in HEM combined with the proposed regression network greatly speeds up the inference while reducing the number of parameters.

From Table V, we can see that while the inference speed of Model1 is significantly improved when only HEM is added, the use of concat instead of cost volume leads to an accuracy loss to some extent. On this occasion, we further added FEM (i.e., Model2). Relative to Model1 and UDISNet, the RMSE of Model2 is improved by 25.2% and 14.4%, respectively. This shows that the FEM is able to adaptively adjust the sampling position to effectively capture nonlinear variations to better accommodate different deformations and structural variability, thus providing more accurate alignment results.

In addition, the RMSE is further optimized by 0.7% than Model2 after using our proposed new edge loss function with no increase in inference time or model parameters. This is due to the fact that the edge loss function focuses on overlapping regions, which enhances the sensitivity to structural differences and ensures that critical edge features receive sufficient attention during model training.

C. About the Restoration Network

To validate the effectiveness of our proposed FSRNet for image stitching restoration, two different methods are compared with ours on the UDIS-D dataset, namely, the direct stacking method (called as direct method) and the restoration method used in UDISNet (called as UDISNet also). Here only UDIS-D is used, because the results are good in the synthetic dataset, which makes it difficult to compare the performance of the models about the restoration network. To show the effects of the different restoration methods, the input of them is the same distorted image pairs generated by our FIANet network. Some examples of the output results by each restoration method are shown in Fig. 9.

The results in Fig. 9 show that the direct stacking solution cannot deal with the ghosting problem caused by imperfectly aligned images. But both our method and UDISNet are able to effectively eliminate ghosting and achieve image fusion with natural transitions (see the part of the tree in the first row of Fig. 9). Compared to the output of UDISNet, our method is visually superior when dealing with ghosting of small moving objects and color transitions in large low-textured areas (see the walking people of the second row in Fig. 9). Furthermore, in the operation of image restoration, the average inference time of UDISNet is about 5.355 s, but our method takes only 1.426 s. These results demonstrate the efficiency and superiority of our proposed FSRNet.

D. About the Real-World Application Performance

The experimental results on the synthetic MS-COCO dataset and the UDIS-D dataset show that the proposed model can get better stitching performance (see Section IV-D). To further test the performance of the proposed model in real-world applications, some experiments are conducted on an aerial dataset [65]. This dataset is randomly captured by drones in several real-world scenarios and includes about 300 high-definition aerial photographs. Aerial photographs pose unique challenges to the image stitching task due to their special

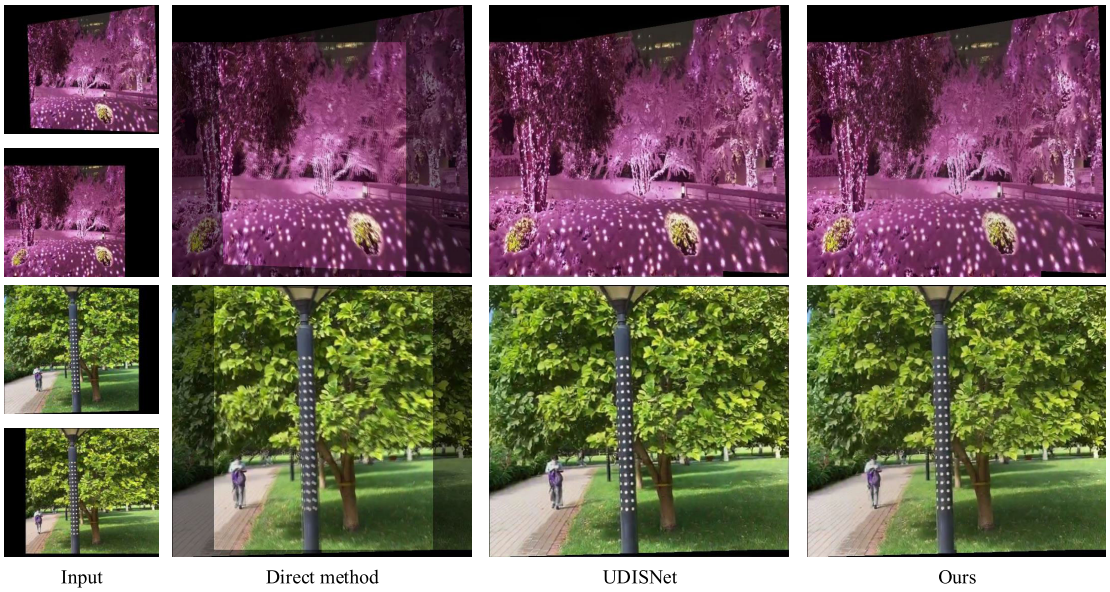


Fig. 9. Some examples of the output results by each restoration method on the UDIS-D dataset.

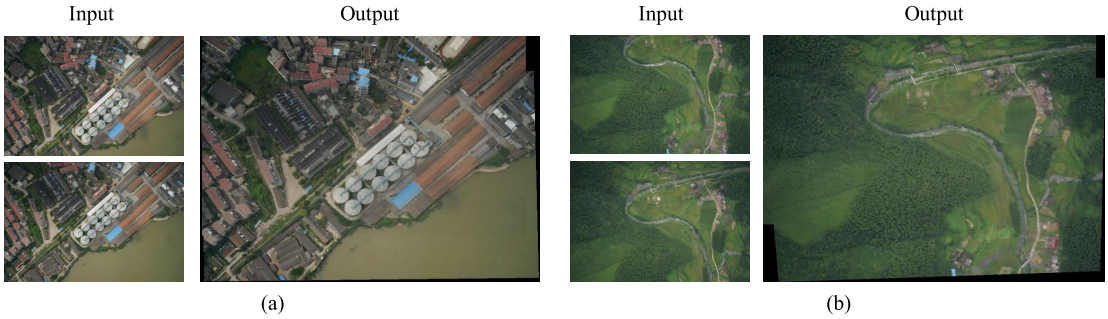


Fig. 10. Stitching result on the aerial dataset based on the proposed model. (a) Multitextured and HR images. (b) Low-textured and LR images.

shooting angle and wide field of view. The results of this experiment are shown in Fig. 10.

As shown in Fig. 10, it can be seen that our method can achieve satisfactory stitching results when dealing with both low-textured as well as multitextured and HR images. In summary, our proposed network has the best stitching effect in handling various scenes, including some challenging real-world applications.

VI. Conclusion

In this article, we investigate the deep learning-based image stitching methods and propose a FUISNet including an improved FIANet and an improved FSRNet. In the proposed FIANet, a modified YOLO-based backbone structure is presented, where an integrated DCM is added. We use a fast HEM to speed up the computational efficiency. In addition, an improved loss function is designed for the image alignment. In the proposed FSRNet, a more compact and efficient reconstruction network with fewer parameters is presented. Various experiments are conducted to test the performance of the proposed model on the homography estimates and the final image stitching tasks. The results show that the proposed model achieves competitive accuracy performance in the image stitching tasks, with very small model size and relatively high speed, compared with the state-of-the-art models.

However, there is a lot of potential room for improvement and optimization in the proposed model. For example, the proposed model mainly focuses on acquiring a single homography matrix through a network model to accomplish image alignment and stitching. When dealing with more complex scenes (such as dynamic scenes with drastic changes in shooting angles and lighting conditions), the sole homography matrix does not express the mapping relationship between images well. Therefore, designing an image stitching model that can handle various complex scenes is a good research direction for the future. In addition, the workflow of the proposed image stitching model is still relatively complicated. The lightweight issue is still not solved well, which leads to the processing speed of the model to be still not good enough. In future work, the workflow should be further optimized and the lightweight model should be further studied, to realize real-time image stitching, which will be useful for more practical applications.

References

[1] M. Fu et al., "Image stitching techniques applied to plane or 3-D models: A review," *IEEE Sensors J.*, vol. 23, no. 8, pp. 8060–8079, Apr. 2023.

[2] T. Liao and N. Li, "Single-perspective warps in natural image stitching," *IEEE Trans. Image Process.*, vol. 29, pp. 724–735, 2020.

[3] J. Du et al., "Fast multispectral fusion and high-precision interdetector image stitching of agile satellites based on velocity vector field," *IEEE Sensors J.*, vol. 22, no. 22, pp. 22134–22147, Nov. 2022.

[4] L. Tian, Q. Li, L. He, and D. Zhang, "Image-range stitching and semantic-based crack detection methods for tunnel inspection vehicles," *Remote Sens.*, vol. 15, no. 21, p. 5158, Oct. 2023.

[5] Y. Fu, T. Guo, and X. Zhao, "Intelligent splicing method of virtual reality Lingnan cultural heritage panorama based on automatic machine learning," *Mobile Inf. Syst.*, vol. 2021, pp. 1–10, Aug. 2021.

[6] J. Ni, Y. Chen, Y. Chen, J. Zhu, D. Ali, and W. Cao, "A survey on theories and applications for self-driving cars based on deep learning methods," *Appl. Sci.*, vol. 10, no. 8, p. 2749, Apr. 2020.

[7] L. Wang, W. Yu, and B. Li, "Multi-scenes image stitching based on autonomous driving," in *Proc. IEEE 4th Inf. Technol., Netw., Electron. Autom. Control Conf. (ITNEC)*, vol. 1, Chongqing, China, Jun. 2020, pp. 694–698.

[8] Y. Yan, L. Xu, Y. Liu, X. Su, J. Gao, and M. Wan, "Larynx ultrasound image stitching based on multiconstraint super-pixel feature," *IEEE Trans. Instrum. Meas.*, vol. 72, 2023, Art. no. 4002311.

[9] J. Wang, Z. Gong, B. Tao, and Z. Yin, "A 3-D reconstruction method for large freeform surfaces based on mobile robotic measurement and global optimization," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 5006809.

[10] J. Ni, T. Gong, Y. Gu, J. Zhu, and X. Fan, "An improved deep residual network-based semantic simultaneous localization and mapping method for monocular vision robot," *Comput. Intell. Neurosci.*, vol. 2020, Feb. 2020, Art. no. 7490840.

[11] Q. Sun, M. Liu, S. Chen, F. Lu, and M. Xing, "Ship detection in SAR images based on multilevel superpixel segmentation and fuzzy fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5206215.

[12] X. Zhu, B. Guo, W. Hu, L. Shi, J. Ma, and D. Xue, "Scene segmentation of multi-band ISAR fusion imaging based on MB-PCSBL," *IEEE Sensors J.*, vol. 21, no. 3, pp. 3520–3532, Feb. 2021.

[13] Z. Wang and Z. Yang, "Review on image-stitching techniques," *Multimedia Syst.*, vol. 26, no. 4, pp. 413–430, Aug. 2020.

[14] J. Wang and Y. Wang, "Modified SURF applied in remote sensing image stitching," *Int. J. Signal Process., Image Process. Pattern Recognit.*, vol. 8, no. 8, pp. 1–10, Aug. 2015.

[15] J. Zhang, G. Chen, and Z. Jia, "An image stitching algorithm based on histogram matching and SIFT algorithm," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 31, no. 4, Apr. 2017, Art. no. 1754006.

[16] C. Sun, X. Wu, J. Sun, N. Qiao, and C. Sun, "Multi-stage refinement feature matching using adaptive ORB features for robotic vision navigation," *IEEE Sensors J.*, vol. 22, no. 3, pp. 2603–2617, Feb. 2022.

[17] A. H. Madessa et al., "Transmittance surface detection and material identification using multitask ViT-SIFT fusion," *IEEE Trans. Instrum. Meas.*, vol. 72, 2023, Art. no. 5003717.

[18] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, Jun. 2008.

[19] Z. Zhou, Y. Zhang, Z. Gu, and S. X. Yang, "Deep learning approaches for object recognition in plant diseases: A review," *Intell. Robot.*, vol. 3, no. 4, pp. 514–537, Oct. 2023.

[20] F. Gan, H. Shao, and B. Xia, "An adaptive model with dual-dimensional attention for remaining useful life prediction of aero-engine," *Knowl.-Based Syst.*, vol. 293, Jun. 2024, Art. no. 111738.

[21] Y. Guo, D. Zhou, P. Li, C. Li, and J. Cao, "Context-aware poly(A) signal prediction model via deep spatial–temporal neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 8241–8253, Jun. 2024.

[22] M. Huber, S. Ourselin, C. Bergeles, and T. Vercauteren, "Deep homography estimation in dynamic surgical scenes for laparoscopic camera motion extraction," *Comput. Methods Biomechanics Biomed. Eng., Imag. Visualizat.*, vol. 10, no. 3, pp. 321–329, May 2022.

[23] T. Nguyen, S. W. Chen, S. S. Shivakumar, C. J. Taylor, and V. Kumar, "Unsupervised deep homography: A fast and robust homography estimation model," *IEEE Robot. Autom. Lett.*, vol. 3, no. 3, pp. 2346–2353, Jul. 2018.

[24] J. Zhang et al., "Content-aware unsupervised deep homography estimation," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K., Aug. 2020, pp. 653–669.

[25] S.-Y. Cao, J. Hu, Z. Sheng, and H.-L. Shen, "Iterative deep homography estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1869–1878.

[26] L. Nie, C. Lin, K. Liao, M. Liu, and Y. Zhao, "A view-free image stitching network based on global homography," *J. Vis. Commun. Image Represent.*, vol. 73, Nov. 2020, Art. no. 102950.

[27] L. Nie, C. Lin, K. Liao, and Y. Zhao, "Learning edge-preserved image stitching from multi-scale deep homography," *Neurocomputing*, vol. 491, pp. 533–543, Jun. 2022.

[28] D.-Y. Song, G. Lee, H. Lee, G.-M. Um, and D. Cho, "Weakly-supervised stitching network for real-world panoramic image generation," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*, Tel Aviv, Israel, Oct. 2022, pp. 54–71.

[29] Z. Shi, H. Li, Q. Cao, H. Ren, and B. Fan, "An image mosaic method based on convolutional neural network semantic features extraction," *J. Signal Process. Syst.*, vol. 92, no. 4, pp. 435–444, Apr. 2020.

[30] L. Nie, C. Lin, K. Liao, S. Liu, and Y. Zhao, "Unsupervised deep image stitching: Reconstructing stitched features to images," *IEEE Trans. Image Process.*, vol. 30, pp. 6184–6197, 2021.

[31] J. Hou and Z. Su, "Research on image stitching algorithm for UAV ground station terminal," in *Proc. Commun. Comput. Inf. Sci.*, vol. 644, Beijing, China, 2016, pp. 207–215.

[32] Z. Li, B. Qi, and C. Niu, "Realtime registration for phased-array submarine sonar image," in *Proc. 10th Int. Conf. Intell. Human-Machine Syst. Cybern. (IHMSC)*, vol. 2, Hangzhou, China, Aug. 2018, pp. 320–324.

[33] M. Brown and D. G. Lowe, "Automatic panoramic image stitching using invariant features," *Int. J. Comput. Vis.*, vol. 74, no. 1, pp. 59–73, Aug. 2007.

[34] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[35] T. Gu, Z. Luo, T. Guo, and T. Luo, "A new reconstruction method for measurement data with multiple outliers," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–9, 2022.

[36] J. Zaragoza, T.-J. Chin, Q.-H. Tran, M. S. Brown, and D. Suter, "As-projective-as-possible image stitching with moving DLT," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1285–1298, Jul. 2014.

[37] K.-Y. Lee and J.-Y. Sim, "Warping residual based image stitching for large parallax," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8195–8203.

[38] Q. Li, Y. Xu, and P. Ding, "PRESCAN adaptive vehicle image real-time stitching algorithm based on improved SIFT," *Int. J. Inf. Technol. Syst. Approach*, vol. 16, no. 3, pp. 1–17, Apr. 2023.

[39] S. Wen, X. Wang, W. Zhang, G. Wang, M. Huang, and B. Yu, "Structure preservation and seam optimization for parallax-tolerant image stitching," *IEEE Access*, vol. 10, pp. 78713–78725, 2022.

[40] C. He, Y. Li, Y. Zhang, and M. Liao, "Region-based seam optimization for image stitching," *J. Electron. Imag.*, vol. 28, no. 4, p. 1, Aug. 2019.

[41] T. Liao, J. Chen, and Y. Xu, "Quality evaluation-based iterative seam estimation for image stitching," *Signal, Image Video Process.*, vol. 13, no. 6, pp. 1199–1206, Sep. 2019.

[42] J. Ni, Y. Chen, G. Tang, J. Shi, W. Cao, and P. Shi, "Deep learning-based scene understanding for autonomous robots: A survey," *Intell. Robot.*, vol. 3, no. 3, pp. 374–401, Aug. 2023.

[43] W. Li, Y. Guo, B. Wang, and B. Yang, "Learning spatiotemporal embedding with gated convolutional recurrent networks for translation initiation site prediction," *Pattern Recognit.*, vol. 136, Apr. 2023, Art. no. 109234.

[44] S. Yang, D. Zhou, J. Cao, and Y. Guo, "LightingNet: An integrated learning method for low-light image enhancement," *IEEE Trans. Comput. Imag.*, vol. 9, pp. 29–42, 2023.

[45] S. Liu et al., "Content-aware unsupervised deep homography estimation and its extensions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 2849–2863, Mar. 2023.

[46] V.-D. Hoang, D.-P. Tran, N. G. Nhu, T.-A. Pham, and V.-H. Pham, "Deep feature extraction for panoramic image stitching," in *Proc. 12th Asian Conf. Intell. Inf. Database Syst. (ACIIDS)*, Phuket, Thailand, Mar. 2020, pp. 141–151.

[47] Q. Zhao, Y. Ma, C. Zhu, C. Yao, B. Feng, and F. Dai, "Image stitching via deep homography estimation," *Neurocomputing*, vol. 450, pp. 219–229, Aug. 2021.

[48] D.-Y. Song, G.-M. Um, H. K. Lee, and D. Cho, "End-to-end image stitching network via multi-homography estimation," *IEEE Signal Process. Lett.*, vol. 28, pp. 763–767, 2021.

[49] H. Kweon, H. Kim, Y. Kang, Y. Yoon, W. Jeong, and K.-J. Yoon, "Pixel-wise warping for deep image stitching," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, vol. 37, no. 1, pp. 1196–1204.

[50] L. Nie, C. Lin, K. Liao, S. Liu, and Y. Zhao, "Parallax-tolerant unsupervised deep image stitching," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 7365–7374.

[51] J. Ni, K. Shen, Y. Chen, W. Cao, and S. X. Yang, "An improved deep network-based scene classification method for self-driving cars," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 5001614.
[52] C. Wang, Z. Wang, K. Li, R. Gao, and L. Yan, "Lightweight object detection model fused with feature pyramid," *Multimedia Tools Appl.*, vol. 82, no. 1, pp. 601–618, Jan. 2023.
[53] J. Ni, K. Shen, Y. Chen, and S. X. Yang, "An improved SSD-like deep network-based object detection method for indoor scenes," *IEEE Trans. Instrum. Meas.*, vol. 72, 2023, Art. no. 5006915.
[54] J. Dai et al., "Deformable convolutional networks," in *Proc. 16th IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 764–773.
[55] S. Xu, L. Zhang, W. Huang, H. Wu, and A. Song, "Deformable convolutional networks for multimodal human activity recognition using wearable sensors," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–14, 2022.
[56] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5410–5418.
[57] H. Sildir, E. Aydin, and T. Kavzoglu, "Design of feedforward neural networks in the classification of hyperspectral imagery using superstructural optimization," *Remote Sens.*, vol. 12, no. 6, p. 956, Mar. 2020.
[58] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. 29th Annu. Conf. Neural Inf. Process. Syst. (NIPS)*, Montreal, QC, Canada, Dec. 2015, pp. 2017–2025.
[59] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Deep image homography estimation," 2016, *arXiv:1606.03798*.
[60] C. Shen, X. Ji, and C. Miao, "Real-time image stitching with convolutional neural networks," in *Proc. IEEE Int. Conf. Real-Time Comput. Robot. (RCAR)*, Irkutsk, Russia, Aug. 2019, pp. 192–197.
[61] H.-F. Wang, Y.-F. Wang, J.-J. Zhang, and J. Cao, "Laser stripe center detection under the condition of uneven scattering metal surface for geometric measurement," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 5, pp. 2182–2192, May 2020.
[62] M. Liu, Y. Chen, J. Xie, L. He, and Y. Zhang, "LF-YOLO: A lighter and faster YOLO for weld defect detection of X-ray image," *IEEE Sensors J.*, vol. 23, no. 7, pp. 7430–7439, Apr. 2023.
[63] H. Wang et al., "Convolutional neural network with a learnable spatial activation function for SAR image despeckling and forest image analysis," *Remote Sens.*, vol. 13, no. 17, p. 3444, Aug. 2021.
[64] M. Brown and D. Lowe, "Recognising panoramas," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Nice, Nice, France, Oct. 2003, pp. 1218–1225.
[65] F. Zhu, J. Li, B. Zhu, H. Li, and G. Liu, "UAV remote sensing image stitching via improved VGG16 Siamese feature extraction network," *Expert Syst. Appl.*, vol. 229, Nov. 2023, Art. no. 120525.

Jianjun Ni (Senior Member, IEEE) received the Ph.D. degree from the School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou, China, in 2005.

He was a Visiting Professor with the University of Guelph, Guelph, ON, Canada, and the University of Essex, Colchester, U.K. He is currently a Professor with the College of Artificial Intelligence and Automation, Hohai University, Changzhou, China. He has published over 100 papers in related international conferences and journals. His research interests include control systems, neural networks, robotics, machine intelligence, and multiagent systems.

Dr. Ni serves as an associate editor and a reviewer for several international journals.

Yingqi Li received the B.S. degree from Nanjing Xiaozhuang University, Nanjing, China, in 2021. She is currently pursuing the M.S. degree with the Department of Electronic Information, College of Information Science and Engineering, Hohai University, Changzhou, China.

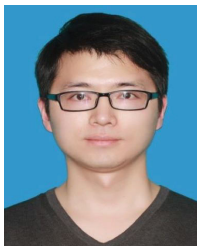
Her research interests include machine learning and image understanding.

Chunyan Ke received the M.S. degree from the Xi'an University of Science and Technology, Xi'an, China, in 2013. She is currently pursuing the Ph.D. degree in IoT technology and application with the College of Information Science and Engineering, Hohai University, Changzhou, China.

Her research interests include computer vision, smart farming technology, and machine learning.

Ziru Zhang received the B.S. degree from Hohai University, Changzhou, China, in 2023, where she is currently pursuing the Ph.D. degree in artificial intelligence with the College of Artificial Intelligence and Automation.

Her research interests include semantic segmentation, target coverage, and robot control.



Weidong Cao (Member, IEEE) received the Ph.D. degree in mechanical engineering from Chongqing University, Chongqing, China, in 2018.

He is currently working as a Lecturer with the College of Artificial Intelligence and Automation, Hohai University, Changzhou, China. His research interests include swarm intelligence optimization algorithms, machine learning, and data-driven modeling.

Simon X. Yang (Senior Member, IEEE) received the B.Sc. degree in engineering physics from Beijing University, Beijing, China, in 1987, the joint M.Sc. degree in biophysics from the Chinese Academy of Sciences, Beijing, in 1990, the M.Sc. degree in electrical engineering from the University of Houston, Houston, TX, USA, in 1996, and the Ph.D. degree in electrical and computer engineering from the University of Alberta, Edmonton, AB, Canada, in 1999.

He is currently a Professor and the Head of the Advanced Robotics and Intelligent Systems Laboratory, University of Guelph, Guelph, ON, Canada. His research interests include robotics, intelligent systems, sensors and multisensor fusion, wireless sensor networks, control systems, transportation, and computational neuroscience.

Dr. Yang serves as the Editor-in-Chief for the *International Journal of Robotics and Automation* and an Associate Editor for IEEE TRANSACTIONS ON CYBERNETICS and several other journals.