Unsupervised Deep Image Stitching: Reconstructing Stitched Features to Images

Lang Nie, Chunyu Lin, *Member, IEEE*, Kang Liao, Shuaicheng Liu, *Member, IEEE*, and Yao Zhao, *Senior Member, IEEE*

***Abstract*— Traditional feature-based image stitching technologies rely heavily on feature detection quality, often failing to stitch images with few features or low resolution. The learning-based image stitching solutions are rarely studied due to the lack of labeled data, making the supervised methods unreliable. To address the above limitations, we propose an unsupervised deep image stitching framework consisting of two stages: unsupervised coarse image alignment and unsupervised image reconstruction. In the first stage, we design an ablation-based loss to constrain an unsupervised homography network, which is more suitable for large-baseline scenes. Moreover, a transformer layer is introduced to warp the input images in the stitching-domain space. In the second stage, motivated by the insight that the misalignments in pixel-level can be eliminated to a certain extent in feature-level, we design an unsupervised image reconstruction network to eliminate the artifacts from features to pixels. Specifically, the reconstruction network can be implemented by a low-resolution deformation branch and a high-resolution refined branch, learning the deformation rules of image stitching and enhancing the resolution simultaneously. To establish an evaluation benchmark and train the learning framework, a comprehensive real-world image dataset for unsupervised deep image stitching is presented and released. Extensive experiments well demonstrate the superiority of our method over other state-of-the-art solutions. Even compared with the supervised solutions, our image stitching quality is still preferred by users.**

***Index Terms*— Computer vision, deep image stitching, deep homogrpahy estimation.**

I. Introduction

> Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.
>
> George E. P. Box

IMAGE stitching is a crucial and challenging computer vision task that has been well-studied in the past decades, with the purpose to construct a panorama with a wider field-of-view (FOV) from different images captured from different viewing positions. This technology can be of great use in varying fields such as biology [1], [2], medical [3], surveillance videos [4], [5], autonomous driving [6], [7], virtual reality (VR) [8], [9].

Conventional image stitching solutions are feature-based methods, where feature detection is the first step that can profoundly affect stitching performance. Then a parametric image alignment model can be established using the matched features, by which we can warp the target image to align with the reference image. Finally, the stitched image can be obtained by assigning pixel values to each pixel in overlapping areas between the warped images.

Among these steps, establishing a parametric image alignment model is crucial in the feature-based methods. In fact, the homography transformation is the most used image alignment model, which contains translation, rotation, scaling, and vanishing point transformation, accounting for the transformation from one 2D plane to another [10] correctly. However, each image domain may contain multiple different depth levels in actual scenes, which contradicts the planar scene assumption of the homography. There are often ghosting effects in the stitched results since a single homography cannot account for all the alignments at different depth levels.

Conventional feature-based solutions alleviate the artifacts in two mainstream ways. The first way is to eliminate the artifacts by aligning the target image with the reference image as much as possible [11]–[20]. These methods partition an image into different areas and compute the homography matrix for each diverse area. By exerting spatially-varying warpings on these areas, the overlapping areas are well aligned, and the artifacts are significantly reduced. The second way is to hide the artifacts by researching for an optimal seam to stitch the warped images [21]–[26]. Through optimizing a seam-related cost, the overlapping can be divided into two complementary regions along the seam. Then, a stitched image is formed according to two regions. The feature-based solutions can significantly reduce the artifacts in most scenes. Still, they rely heavily on feature detection so that the stitching performance can drop sharply or even fail in scenes with few features or at low resolution.

Due to the incredible feature extraction capability of Convolutional Neural Networks (CNNs), recently learning-based approaches have achieved state-of-the-art performance in various fields such as depth estimation [28], optical flow estimation [29], [30], distortion rectification [31]. Increasing researchers try to apply CNNs to image stitching. In [32], [33], the CNNs are only used to extract feature points, while

Manuscript received January 24, 2021; revised May 29, 2021 and June 22, 2021; accepted June 23, 2021. Date of publication July 2, 2021; date of current version July 9, 2021. This work was supported by the National Natural Science Foundation of China under Grant 61772066 and Grant 61972028. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Nikolaos Mitianoudis. *(Corresponding author: Chunyu Lin.)*

Lang Nie, Chunyu Lin, Kang Liao, and Yao Zhao are with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China, and also with the Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China (e-mail: nielang@bjtu.edu.cn; cylin@bjtu.edu.cn; kang_liao@bjtu.edu.cn; yzhao@bjtu.edu.cn).

Shuaicheng Liu is with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: liushuaicheng@uestc.edu.cn).

Data is available on-line at www.github.com/nie-lang/UnsupervisedDeepImageStitching.

Digital Object Identifier 10.1109/TIP.2021.3092828

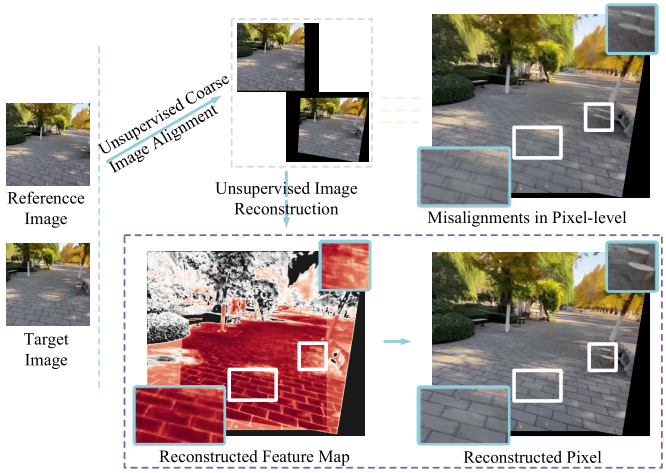


Fig. 1. The pipeline of proposed unsupervised deep image stitching. In the coarse alignment stage, the inputs are warped using a single homography. In the reconstruction stage, the warped images are used for reconstructing the stitched image from feature to pixel.

in [4], [7], [34], the CNNs are proposed to stitch images with fixed viewing positions. Regrettably, these methods are either not a complete learning-based framework [32], [33], or can only be used to stitch images with fixed views instead of arbitrary views [4], [7], [34]. Then, view-free deep image stitching methods [35], [36] are proposed to overcome the two problems simultaneously. In these view-free solutions, deep image stitching can be completed by a deep homography module, a spatial transformer module, and a deep image refined module. However, all the solutions are supervised methods, and there is no real dataset for deep image stitching because of the unavailability of stitched labels in actual scenes until now. Therefore, these networks can only be trained on a 'no-parallax' synthetic dataset, resulting in unsatisfying applications in real scenes.

To overcome the limitations of feature-based solutions and supervised deep solutions, we propose an unsupervised deep image stitching framework that comprises an unsupervised coarse image alignment stage and an unsupervised image reconstruction stage. The pipeline is shown in Fig. 1. In the first stage, we coarsely align the input images using a single homography. Different from the existing unsupervised deep homography solutions [37], [38] that require extra image contents around the input images as supervision, we design an ablation-based loss to optimize our unsupervised deep homography network that is more suitable for the large-baseline scenes, where large-baseline is a relative concept to small-baseline in [38]. Besides, a stitching-domain transformer layer is proposed to warp the input images in the stitching-domain with less occupied space than the existing deep stitching works [35], [36]. In the second stage, we present an ingenious strategy to reconstruct the stitched images from feature to pixel, eliminating the artifacts by unsupervised image reconstruction. In particular, we design a low-resolution deformation branch and a high-resolution refined branch in the reconstruction network to learn the deformation rules of image stitching and enhances the resolution, respectively.

This reconstruction strategy is motivated by an observation: misalignments in feature-level are more unnoticeable than

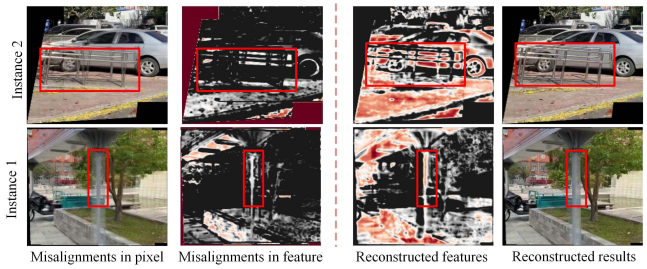


Fig. 2. Motivation: the misalignments in pixel-level can be visually weakened in feature-level. Col 1: the results of stitching the warped images from unsupervised coarse alignment stage. Col 2: the results of stitching the warped features extracted by the 'conv1_2' in VGG19 [27]. Col 3-4: reconstructing from feature to pixel by unsupervised reconstruction network.

in pixel-level (Fig. 2 left). Compared with pixels, feature maps are more blurred, which indicates the misalignments in pixel-level can be eliminated to a certain extent in feature-level. Therefore, we believe it is easier to eliminate artifacts in feature-level than in pixel-level. To implement this, we first reconstruct the features of the stitched image that are as close to the two warped images as possible (Col 3 in Fig. 2). Then the stitched image can then be reconstructed at pixel-level (Col 4 in Fig. 2) based on the reconstructed features.

The existing dataset in learning-based solutions [35], [36] is a 'no-parallax' synthetic dataset that cannot represent the practical application scene. And the datasets in feature-based solutions are too few to support deep learning training. To enable our framework the generalization ability in real scenarios, we also propose a large real-world image stitching dataset containing varying overlap rates, varying degrees of parallax, and variable scenes such as indoor, outdoor, night, dark, snow, and zooming. Here, we define overlap rate as the percentage of the overlapping area in the total area of the image.

In experiments, we evaluate our performance in homography estimation and image stitching. Experimental results demonstrate the superiority of our method over other state-of-the-art solutions in real scenes. The contributions of this paper are summarized as follows:

- We present an unsupervised deep image stitching framework consisting of an unsupervised coarse image alignment stage and an unsupervised image reconstruction stage.
- We propose the first large real dataset for unsupervised deep image stitching (to the best of our knowledge), which we hope can work as a benchmark dataset and promote other related research work.
- Our algorithm outperforms the state-of-the-art, including homography estimation solutions and image stitching solutions in real scenes. Even compared with the supervised solutions, our image stitching quality is still preferred by users.

II. Related Work

In this section, we subsequently review the existing works in image stitching and deep homography estimation.

A. Feature-Based Image Stitching

According to different strategies to eliminate artifacts, the feature-based image stitching algorithms can be divided into the following two categories:

1) Adaptive Warping Methods: Considering that a single transformation model is not enough to accurately align images with parallax, the idea of combining multiple parametric alignment models to align the images as much as possible is introduced. In [11], the dual-homography warping (DHW) is presented to align the foreground and the background, respectively. This method works well in the scene composed of two predominating planes but shows poor performance in more complex scenes. Lin *et al.* [12] apply multiple smoothly varying affine (SVA) transformations in different regions, enhancing local deformation and alignment performance. Zaragoza *et al.* [13] propose the as-projective-as-possible (APAP) approach, where an image can be partitioned into dense grids, and each grid would be allocated a corresponding homography by weighting the features. In fact, APAP would still exhibit parallax artifacts in the vicinity of the object boundaries, for dramatic depth changes might occur in these areas. To get rid of this problem, the warping residual vectors are proposed to distinguish matching features from different depth planes in [19], contributing to more naturally stitched images.

2) Seam-Driven Methods: Seam-driven image stitching methods are also influential, acquiring natural stitched images by hiding the artifacts. Inspired by the idea of interactive digital photomontage [39], Gao *et al.* [24] propose to choose the best homography with the lowest seam-related cost from candidate homography matrices. Then the artifacts are hidden through seam cutting. Referring to the optimization strategy of content-preserving warps (CPW) [40], Zhang and Liu [22] propose a seam-based local alignment approach while maintaining the global image structure using an optimal homography. This work was also extended to stereoscopic image stitching [41]. Using the iterative warp and seam estimation, Lin *et al.* [23] find the optimal local area to stitch images, which can protect the curve and line structure during image stitching.

These feature-based algorithms contribute to perceptually nature stitched results. However, they rely heavily on the quality of feature detection, often failing in scenes with few features or at low resolution.

B. Learning-Based Image Stitching

Getting a real dataset for stitching is difficult. In addition, deep stitching is quite challenging for the scenes with low overlap rate and large parallax. Subjected to these two problems, learning-based image stitching is still in development.

1) View-Fixed Methods: View-fixed image stitching methods are task-driven, which are designed for the specific application scenes such as autonomous driving [6], [7], surveillance videos [4]. In these works, the end-to-end networks are proposed to stitch images from fixed views while they cannot be extended to stitch images from arbitrary views.

2) View-Free Methods: To stitch images from arbitrary views using CNNs, some researchers propose to adopt CNNs in the stage of feature detection [32], [33]. However, these methods cannot be regarded as a complete learning-based framework strictly. The first complete learning-based framework to stitch images from arbitrary views was proposed in [35]. The images can be stitched through three stages: homography estimation, spatial transformation, and content refinement. Nevertheless, this work cannot handle input images with arbitrary resolutions due to the fully connected layers in the network, and the stitching quality in real applications is unsatisfying. Following this deep stitching pipeline, an edge-preserved deep image stitching solution was proposed in [36], freeing the limitation of input resolution and significantly improving the stitching performance in real scenes.

C. Deep Homography Schemes

The first deep homography method was put forward in [42], where a VGG-style [27] network was used to predict the eight offsets of four vertices of an image, thus uniquely determine a corresponding homography. Nguyen *et al.* [37] proposed the first unsupervised deep homography approach with the same architecture as [42] with an effective unsupervised loss. Introducing spatial attention to deep homography network, Zhang *et al.* [38] proposes a content-aware unsupervised network, contributing to SOTA performance in small-baseline deep homography. In [43], multi-scale features are extracted to predict the homography from coarse to fine using image pyramids.

Besides that, the deep homography network is usually adopted as a part of the view-free image stitching frameworks [35], [36]. Different from [37], [38], [42], [43], the deep homography in image stitching is more challenging, for the baseline between input images is usually 2X∼3X larger.

III. Unsupervised Coarse Image Alignment

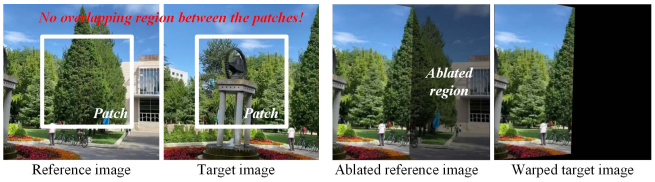
Given two high-resolution input images, we first estimate the homography using a deep homography network in an unsupervised manner. Then the input images can be warped to align each other coarsely in the proposed stitching-domain transformer layer.

A. Unsupervised Homography

The existing unsupervised deep homography methods [37], [38] take the image patches as the input, which is shown in the white squares in Fig. 3(a). The objective function of these methods can be expressed as Eq. (1):

$$L_{PW} = \left\| \mathcal{P}(I^A) - \mathcal{P}(\mathcal{H}(I^B)) \right\|_1, \tag{1}$$

where I^A, I^B represent the full images of the reference image and the target image, respectively. $\mathcal{P}(\cdot)$ is the operation of extracting an image patch from a full image, and $\mathcal{H}(\cdot)$ warps one image to align with the other using estimated homography. From Eq. (1), we can see that to make the warped target patch close to the reference patch, the extra contents around the target patch are utilized to pad the invalid pixels in the warped target patch. We call it a padding-based constraint strategy. This strategy works well in small-baseline [38],



(a) A failure case of padding-based strategy. (b) The proposed ablation-based strategy.

Fig. 3. An instance to show that the proposed ablation-based strategy is more suitable for large-baseline unsupervised homography estimation.

or middle-baseline [37] homography estimations while it fails in the large-baseline case. In particular, when the baseline is too large (as illustrated in Fig. 3(a)), there might be no overlapping area between the input patches, which leads to the meaningless estimation of homography from these patches.

To solve this problem, we design an ablation-based strategy to constrain large-baseline unsupervised homography estimation. Specifically, we take the full images as the input, ensuring that all overlapping areas are included in our inputs. When we enforce the warped target image close to the reference image, we no longer pad the invalid pixels in the warped image. Instead, we ablate the contents in the reference image where the invalid pixels in the warped target image locate, as shown in Fig. 3(b). Our objective function for unsupervised homography is formulated as Eq. (2):

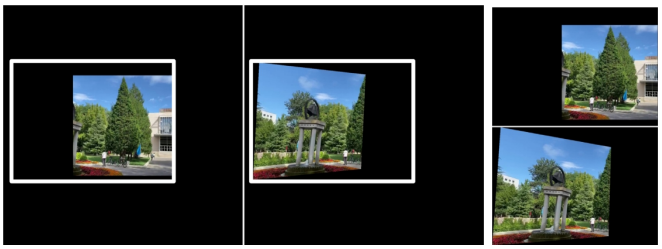
$$L'_{pW} = \left\| \mathcal{H}(E) \odot I^A - \mathcal{H}(I^B) \right\|_1, \tag{2}$$

where $\odot$ is the pixel-wise multiplication and E is an all-one matrix with identical size with I^A.

As for the architecture of our unsupervised homography network, we adopt a multi-scale deep model proposed in [36], which connects feature pyramid and feature correlation in a unified framework so that it can predict the honography from coarse to fine and handle relative large-baseline scenes.

B. Stitching-Domain Transformer Layer

The spatial transformer layer was first proposed in [44], where images can be spatially transformed with gradient backpropagation guaranteed using the homography model. In image stitching, input images of the same resolution can output stitched images of different resolution according to the varying overlapping rates, which brings a considerable challenge to deep image stitching. The existing deep image stitching methods solve this problem by extending the spatial transformer layer [35], [36]. Specifically, these solutions define a maximum resolution for the stitched image so that all the input contents can be included in the output. In addition, the network will output images with the same resolution every time. However, most of the space occupied by black pixels outside the white box in Fig. 4(a) are wasted. To deal with spatial waste, we propose a stitching-domain transformer layer. We define the stitching-domain as the smallest bounding rectangle of the stitched image, which saves the most space while ensuring the integrity of the image contents. The warped results of ours are illustrated in Fig. 4(b), and our stitching-domain transformer layer can be implemented as follows.



(a) (b)

Fig. 4. The comparison between the spatial transformer layer in existing deep image stitching and our stitching-domain transformer layer. (a): Warping by spatial transformer layer in existing deep image stitching [35], [36]. (b): Warping by our stitching-domain transformer layer.

First, we calculate the coordinates of the 4 vertices in the warped target image by Eq. (3):

$$(x_k^W, y_k^W) = (x_k^B, y_k^B) + (\Delta x_k, \Delta y_k), k \in \{1, 2, 3, 4\}, \tag{3}$$

where (x_k^W, y_k^W), (x_k^B, y_k^B) are the k-th vertex coordinates of the warped target image and the target image, respectively. $(\Delta x_k, \Delta y_k)$ donate the offsets of the k-th vertex that are estimated form the aforementioned homogrpahy network. Then, the size of the warped image $(H^* \times W^*)$ can be obtained by Eq. (4):

$$\begin{aligned} W^* &= \max_{k\in\{1,2,3,4\}} \{x_k^W, x_k^A\} - \min_{k\in\{1,2,3,4\}} \{x_k^W, x_k^A\}, \\ H^* &= \max_{k\in\{1,2,3,4\}} \{y_k^W, y_k^A\} - \min_{k\in\{1,2,3,4\}} \{y_k^W, y_k^A\}, \end{aligned} \tag{4}$$

where (x_k^A, y_k^A) are the vertex coordinates of the reference image that have the same values as (x_k^B, y_k^B). Finally, we assign the specific values to the pixels of the warped images (I^{AW}, I^{BW}) from the input images (I^A, I^B), which can be represented as Eq. (5):

$$\begin{aligned} I^{AW} &= \mathcal{W}(I^A, I), \\ I^{BW} &= \mathcal{W}(I^B, H), \end{aligned} \tag{5}$$

where I and H are the identity matrix and the estimated homography matrix, respectively. And $\mathcal{W}(\cdot)$ donates the operation of warping an image using a 3×3 transformation matrix with the stitching-domain set to $H^* \times W^*$.

In this way, we transform the input images in the stitching-domain space, effectively reducing the space occupied by feature maps in the subsequent reconstruction network. Compared with the transformer layer used in [35], [36], the proposed layer can help to stitch larger resolution images when the GPU memory is limited.

IV. Unsupervised Image Reconstruction

Considering the limitation that a single homography can only represent the spatial transformation in the same depth [10], the input images cannot be completely aligned in the real-world dataset in the first stage. To break the bottleneck of single homography, we propose to reconstruct the stitched image from feature to pixel. The overview of the proposed unsupervised deep image stitching framework is illustrated in Fig. 5. The reconstruction network can be implemented by two branches: low-resolution deformation branch (Fig. 5 top) and high-resolution refined branch (Fig. 5 bottom), learning

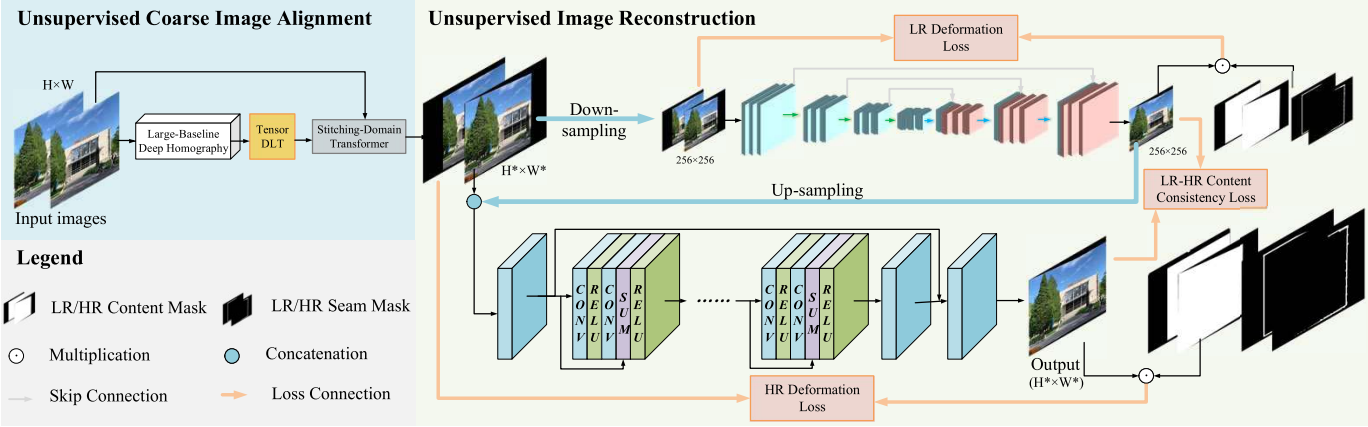


Fig. 5. An overview of our unsupervised deep image stitching. Left: the unsupervised coarse image alignment stage. Right: the unsupervised image reconstruction stage.

the deformation rules of image stitching and enhancing the resolution, respectively.

A. Low-Resolution Deformation Branch

Reconstructing the images only in the high-resolution branch is not appropriate because the receptive field decreases relatively as the resolution increases. To ensure that the receptive field of the network can completely perceive misaligned regions (especially in the case of high resolution and large parallax), we designed a low-resolution branch to learn the deformation rules of image stitching first. As shown in Fig. 5(top), the warped images are first down-sampled to a low-resolution, defined as 256×256, in our implementation. Then an encoder-decoder network consisting of 3 pooling layers and 3 deconvolutional layers is used to reconstruct the stitched image. The filter numbers of the convolutional layers are set to 64, 64, 128, 128, 256, 256, 512, 512, 256, 256, 128, 128, 64, 64, and 3, respectively. Furthermore, skip connections are adopted to connect the low-level and high-level features with the same resolution [45].

In this process, the deformation rules of image stitching are learned with content masks and seam masks (Fig. 6). The content masks are adopted to constrain the features of the reconstructed image close to the warped images, while the seam masks are designed to constrain the edges of the overlapping areas to be natural and continuous. In particular, we obtain the content masks (M^{AC}, M^{BC}) using Eq. (5) by replacing the I^A, I^B with an all-one matrix $E_{H\times W}$, and the seam masks can be calculated by Eq. (6) and Eq. (7):

$$\begin{aligned} \nabla M^{AC} &= |M^{AC}_{i,j} - M^{AC}_{i-1,j}| + |M^{AC}_{i,j} - M^{AC}_{i,j-1}|, \\ \nabla M^{BC} &= |M^{BC}_{i,j} - M^{BC}_{i-1,j}| + |M^{BC}_{i,j} - M^{BC}_{i,j-1}|, \end{aligned} \tag{6}$$

$$\begin{aligned} M^{AS} &= \mathcal{C}(\nabla M^{BC} * E_{3\times3} * E_{3\times3} * E_{3\times3}) \odot M^{AC}, \\ M^{BS} &= \mathcal{C}(\nabla M^{AC} * E_{3\times3} * E_{3\times3} * E_{3\times3}) \odot M^{BC}, \end{aligned} \tag{7}$$

where (i, j) donates the coordinate location, $*$ represents the operation of convolution, and $\mathcal{C}$ clips all the elements to between 0 and 1. Then we design the content loss and seam loss in low-resolution as Eq. (8) and Eq. (9):

$$\begin{aligned} \mathcal{L}^l_{Content} &= \mathcal{L}_P(S_{LR} \odot M^{AC}, I^{AW}) \\ &\quad + \mathcal{L}_P(S_{LR} \odot M^{BC}, I^{BW}), \end{aligned} \tag{8}$$

$$\begin{aligned} \mathcal{L}^l_{Seam} &= \mathcal{L}_1(S_{LR} \odot M^{AS}, I^{AW} \odot M^{AS}) \\ &\quad + \mathcal{L}_1(S_{LR} \odot M^{BS}, I^{BW} \odot M^{BS}) \end{aligned} \tag{9}$$

where S_{LR} is the low-resolution stitched image. $\mathcal{L}_1$ and $\mathcal{L}_P$ donate the L1 loss and the perceptual loss [46], respectively. To make the feature of the reconstructed image as close to that of the warped images as possible, we calculate the perceptual loss on layer 'conv5_3' of VGG-19 [27] which is deep enough to shrink the feature difference between the warped images. Next, the total loss function of low-resolution unsupervised deformation can be formulated as Eq. (10):

$$\mathcal{L}_{LR} = \lambda_c \mathcal{L}^l_{Content} + \lambda_s \mathcal{L}^l_{Seam} \tag{10}$$

where λ_s and λ_c weight the contribution of the content constraint and seam constraint.

B. High-Resolution Refined Branch

After the initialized deformation in the low-resolution branch, we develop a high-resolution refined branch to enhance the resolution and refine the stitched image. The high-resolution refers to the resolution of the output of the first stage. Actually, in our dataset, the resolution is bigger than 512×512. To illustrate the effect of high-resolution branch, we exhibit the outputs of two branches in Fig. 7. This branch is composed of convolutional layers entirely, as shown in Fig. 5 (bottom), which means it can deal with pictures of arbitrary resolution. To be specific, it consists of three separate convolutional layers and eight resblocks [47], of which the filter number of each layer is set to 64 except that of the last layer is set to 3. To prevent low-level information from being gradually forgotten as the convolutional network gets deep, the feature of the first layer is added with that of the penultimate layer. Moreover, each resblock is composed of convolution, relu, convolution, sum, and relu.

We up-sample S_{LR} to the resolution of the warped images and concatenate them together as the input of this branch. The output is the high-resolution stitched image S_{HR}.

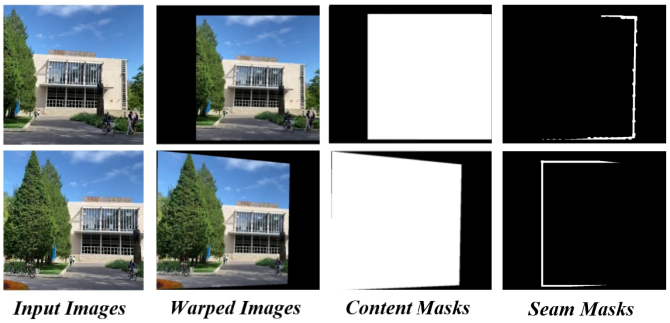


Fig. 6. Learning deformation rules with masks in low-resolution. From left to right, each column represents input images (I^A, I^B), low-resolution warped images (I^{AW}, I^{BW}), content masks (M^{AC}, M^{BC}), and seam masks (M^{AS}, M^{BS}).

Fig. 7. The outputs of the low-resolution branch and high-resolution branch. The high-resolution branch is designed to enhance the resolution and refine the stitched image.

And we conclude the loss function of the high-resolution refined branch $\mathcal{L}_{HR}$ imitating Eq. (10) as Eq. (11):

$$\mathcal{L}_{HR} = \lambda_c \mathcal{L}^h_{Content} + \lambda_s \mathcal{L}^h_{Seam} \tag{11}$$

where $\mathcal{L}^h_{Content}$ and $\mathcal{L}^h_{Seam}$ are the content loss and seam loss in high-resolution which can be calculated using Eq. (8), (9) by replacing the S_{LR} and low-resolution masks with the S_{HR} and the high-resolution masks. When calculating the $\mathcal{L}_P$ in high resolution, we adopt the layer 'conv3_3' of VGG-19, since this layer is shallower than the layer 'conv5_3' (used in $\mathcal{L}_P$ of low resolution) and the output using this layer is more clear.

C. Objective Function

The high-resolution branch is designed to refine the stitched image, but it tends to cause artifacts in the stitched image, since the increase in resolution can relatively reduce the receptive field of the network (more details can be found in Section V-D). To enable our network the abilities to enhance resolution and to eliminate parallax artifacts simultaneously, a content consistency loss is proposed as Eq. (12):

$$\mathcal{L}_{CS} = \left\| S^{256\times256}_{HR} - S_{LR} \right\|_1, \tag{12}$$

where $S^{256\times256}_{HR}$ is obtained by resizing S_{HR} to 256×256 that is the resolution of the output in low-resolution branch.

Taking all the constraints into consideration, we conclude our objective function of the image reconstruction stage as Eq. (13):

$$\mathcal{L}_R = \omega_{LR}\mathcal{L}_{LR} + \omega_{HR}\mathcal{L}_{HR} + \omega_{CS}\mathcal{L}_{CS}, \tag{13}$$

where the ω_{LR}, ω_{HR} and ω_{CS} represent weights of each part.

D. Reconstruction From Feature to Pixel

To exhibit the learning process from feature to pixel, we visualized the feature maps of the low-resolution deformation branch in Fig. 8. At the very beginning of the encoder stage, the network only focuses on the overlapping areas, and the features of non-overlapping areas are all suppressed. Next, as the resolution decreases, deeper semantic features are extracted and reconstructed. In the decoder stage, the network begins to pay attention to non-overlapping areas besides overlapping areas. As the resolution is restored, clearer feature maps are reconstructed. Finally, the stitched image is reconstructed at the pixel level.

V. Experiments

In this section, extensive experiments are conducted to validate effectiveness of the proposed method.

A. Dataset and Implement Details

1) Dataset: To train our network, we also propose an unsupervised deep image stitching dataset that is obtained from variable moving videos. Of these videos, some are from [38] and the others are captured by ourselves. By extracting the frames from these videos with different interval time, we get the samples with different overlap rates (Fig. 9(b)). Moreover, these videos are not captured by the camera rotating around the optical center, and the shot scenes are far from a planar structure, which means this dataset contains different degrees of parallax (Fig. 9(c)). Besides, this real-world dataset includes variable scenes such as indoor, outdoor, night, dark, snow, and zooming (Fig. 9(a)).

To quantitatively describe the distribution of different overlap rates and varying degrees of parallax in our dataset. We divide the overlap rates into 3 levels and define a high overlap rate greater than 90%, a middle overlap rate ranging from 60%-90%, and a low overlap rate lower than 60%. This classification criterion is formulated according to [37], [38], [42], where [38] is the represnetative work in high overlap rate. The average overlap rate of the proposed dataset is greater than 90%. And [37], [42] are the representative works in middle overlap rate for the average overlap rate of Warped COCO (disturbance < 32) dataset [42] is about 75%. Besides, to describe parallax accurately, we align the target image with the reference image using a global homography and then calculate the maximum misalignment error of corresponding feature points in the coarse aligned images to show the magnitude of parallax. In this way, we divide the parallax into 2 levels: small parallax with error smaller than 30 pixels and large parallax with error greater than 30 pixels. Fig. 9(c) demonstrates the difference of different parallax intuitively.

In particular, we get 10,440 cases for training and 1,106 for testing. Among our dataset, the ratios of overlap rates from high to low are about 16%, 66%, and 18%, while the ratios of

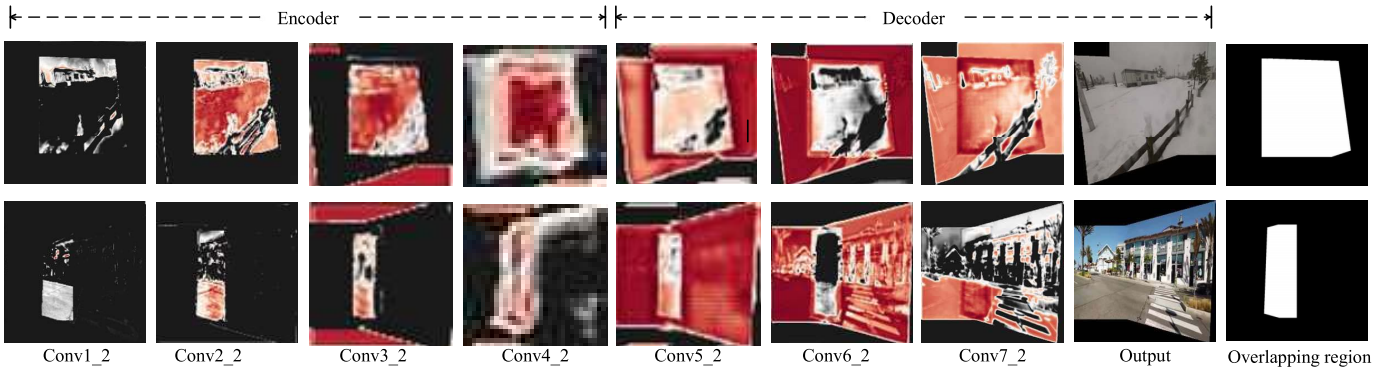
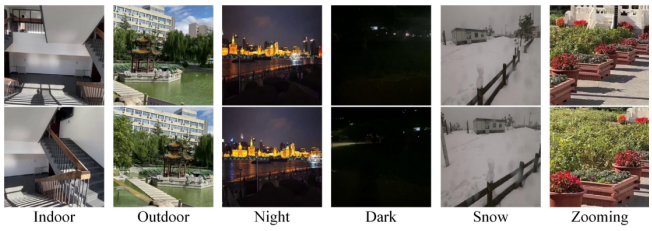
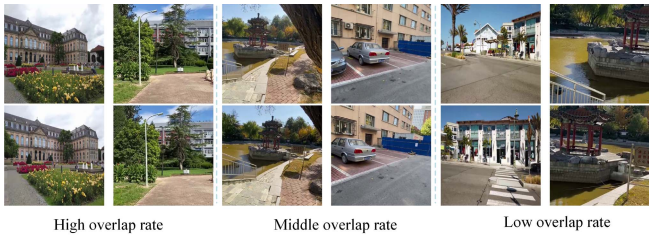


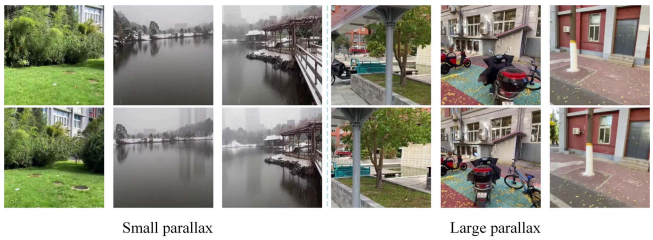
Fig. 8. Visualization of the learning process of the low-resolution deformation branch. The stitched images are reconstructed from overlapping regions to non-overlapping regions.



(a) Varying scenes in our dataset.



(b) Varying overlap rates in our dataset.



(c) Varying degrees of parallax in our dataset.

Fig. 9. Illustrations of our proposed unsupervised deep image stitching dataset.

parallax from small to large are about 91% and 9%. Although our dataset contains no ground-truth, we include our testing results in this dataset, which we hope can work as a benchmark dataset for other researchers to follow and compare.

2) Details: We train our unsupervised image stitching framework in three steps. First, we train our deep homography network on the synthetic dataset (Stitched MS-COCO [35]) for 150 epochs. Second, we finetune the homography network on the proposed real dataset for 50 epochs. Third, we train the deep image reconstruction network on the proposed real dataset for 20 epochs. All the training process is unsupervised, which means our framework only takes the reference/target image as input and requires no label. The optimizer is Adam [48] with an exponentially decaying learning rate with an initial value of 10^{-4}. We set λ_s and λ_c to 2 and 10^{-6}. And ω_{LR}, ω_{HR} and ω_{CS} are set to 100, 1 and 1, respectively. In testing, it takes about $0.4s$ to stitch 2 input images with resolution of 512×512. All the components of this framework are implemented on TensorFlow. Both the training and testing are conducted on a single GPU with NVIDIA RTX 2080 Ti.

B. Comparison of Homography Estimation

To evaluate the performance of the proposed ablation-based unsupervised deep homography objectively, we compare our solution with $I_{3\times3}$, SIFT [49]+RANSAC [50], DHN [42], UDHN [37], CA-UDHN [38], and LB-DHN [36] on the synthetic dataset and real dataset respectively. The $I_{3\times3}$ refers to a 3×3 identity matrix as a 'no-warping' homography for reference, and SIFT+RANSAC is chosen as the representative of traditional homography solutions because it outperforms most traditional solutions as shown in [37], [38]. The DHN, UDHN, CA-UDHN, and LB-DHN are the deep learning solutions, of which UDHN and CA-UDHN are the unsupervised solutions that both adopt the padding-based strategy to train their networks.

1) Synthetic Dataset: The first comparative experiment is conducted on Warped MS-COCO that is the most known synthetic dataset for deep homography estimation. All the learning methods are trained on Warped MS-COCO. The results are illustrated in Table I, where 'Ours_v1' is our model trained with this dataset in an unsupervised manner. From Table I, we can observe:

(1) Ours_v1 outperforms the existing unsupervised deep homography methods (UDHN, CA-UDHN), of which CA-UDHN is the SOTA solution in small-baseline deep homography. However, the performance of CA-UDHN in this dataset degenerates to be close to that of $I_{3\times3}$ due to its limited receptive field.

(2) After adopting our ablation-based unsupervised loss to LB-DHN, 4pt-Homography RMSE increases, which means this loss is not suitable for this 'no-parallax' synthetic dataset.

2) Real Dataset: Then, we carry on a comparison on the proposed real dataset, which consists of varying degrees of parallax. Since this dataset lacks ground truth, we adopt the PSNR and SSIM of the overlapping regions to evaluate the

TABLE I

COMPARISON EXPERIMENT ON HOMOGRAPHY ESTIMATION. THE 1ST AND 2ND BEST SOLUTIONS ARE MARKED IN RED AND BLUE, RESPECTIVELY

(a) 4pt-Homography RMSE (↓) on Warped MS-COCO (synthetic)

Method	Traditional homography		Deep homography (supervised)		Deep homography (unsupervised)		
	$I_{3\times3}$	SIFT [49]+RANSAC [50]	DHN [42]	LB-DHN [36]	UDHN [37]	CA-UDHN [38]	Ours_v1 (synthetic)
Top 0∼30%	15.0154	**0.6743**	3.2998	**0.2719**	2.1894	15.0082	1.1773
30∼60%	18.2515	**1.0964**	4.8839	**0.4140**	3.5272	18.2498	1.4544
60∼100%	21.3517	19.0286	7.6944	**0.9632**	6.4984	21.3618	**3.0702**
Average	18.5220	9.4782	5.5358	**0.5962**	4.3179	18.5234	**2.0239**

(b) PSNR (↑) of the overlapping regions on the proposed dataset (real)

Method	Traditional homography		Deep homography (supervised)		Deep homography (unsupervised)		
	$I_{3\times3}$	SIFT [49]+RANSAC [50]	DHN [42]	LB-DHN [36]	UDHN [37]	Ours_v1 (synthetic)	Ours_v2 (real)
Top 0∼30%	16.1923	25.2300	16.3957	24.7515	19.3851	**26.1958**	**27.8386**
30∼60%	13.0546	22.2308	13.3648	21.1436	15.9251	**22.6115**	**23.9451**
60∼100%	10.8747	17.5791	11.5001	18.4594	13.1016	**19.5277**	**20.7013**
Average	13.1151	21.2541	13.5191	21.1418	15.8252	**22.4421**	**23.8045**

(c) SSIM (↑) of the overlapping regions on the proposed dataset (real)

Method	Traditional homography		Deep homography (supervised)		Deep homography (unsupervised)		
	$I_{3\times3}$	SIFT [49]+RANSAC [50]	DHN [42]	LB-DHN [36]	UDHN [37]	Ours_v1 (synthetic)	Ours_v2 (real)
Top 0∼30%	0.3869	0.8598	0.4088	0.8249	0.5732	**0.8671**	**0.9023**
30∼60%	0.1730	0.7662	0.1699	0.7124	0.3344	**0.7844**	**0.8298**
60∼100%	0.0732	0.5583	0.0772	0.5497	0.1651	**0.6270**	**0.6846**
Average	0.1969	0.7105	0.2042	0.6805	0.3379	**0.7456**	**0.7929**

performance, which can be calculated as Eq. (14):

$$\begin{aligned} PSNR_{overlap} &= \mathcal{PSNR}(\mathcal{H}(E) \odot I^A, \mathcal{H}(I^B)), \\ SSIM_{overlap} &= \mathcal{SSIM}(\mathcal{H}(E) \odot I^A, \mathcal{H}(I^B)), \end{aligned} \tag{14}$$

where $\mathcal{PSNR}(\cdot)$ and $\mathcal{SSIM}(\cdot)$ donates the operations of computing PSNR and SSIM between two images, respectively. We test DHN and UDHN using the public pretrained models. LB-DHN and Ours_v1 are trained on Stitched MS-COCO [35] which is similar to Warped MS-COCO with lower overlap rate. Ours_v2 is the model of finetuning Ours_v1 about 50 epochs on the proposed real dataset. By analyzing the results shown in Table I (b) I(c), we can conclude:

(1) The proposed unsupervised solution (Ours_v2) outperforms all the methods, including the supervised ones in the real dataset.

(2) Although Ours_v1 and LB-DHN are both trained on the synthetic dataset, Ours_v1 achieves better performance under the real dataset, which indicates the proposed unsupervised loss can equip the network with better generalization ability.

C. Comparison of Image Stitching

To verify our method's superiority in image stitching, we compare our method with feature-based solutions and compare with recent learning-based solutions (even if it is not fair to compare our unsupervised algorithms with the supervised ones).

1) Compared with Feature-Based Solutions: In this section, we choose global Homography [10], APAP [13], robust ELA [18] as the representatives of feature-based solutions to compare with our algorithms. Of these methods, we implement Homography with global projective transformation, and we get the stitched results of APAP and robust ELA (adaptive warping methods) by running their open-source codes with our testing instances. After alignment, image fusion is adopted to produce the stitched image and reduce artifacts. Specifically, we fuse the warped images with the pixel-weighted principle, assigning a relatively large weight to the pixel with a high intensity value.

a) Study on Robustness: The performance of feature-based solutions is easily affected by the quantity and distribution of the feature points, resulting in weak robustness in varying scenes. By contrast, the proposed method overcomes this problem. To validate this view, we test the feature-based methods and ours on our test set (1,106 samples). To simulation the change of feature quantity, we resize the test set to different resolutions, e.g., 512×512, 256×256, and 128×128. As the resolution decreases, the number of features decreases exponentially. The results are shown in Table II, where 'error' indicates the number of program crashes and 'failure' refers to the number of stitching unsuccessfully. Specifically, we define significant distortion (Fig. 10 top) and intolerable artifacts (Fig. 10 bottom) as 'failure'. All the stitched results of these methods will be public with our dataset. Comparing the success rates in Table II, we can observe:

(1) Ours is more robust than the feature-based methods. In fact, the 'error' and 'failure' cases of the feature-based solutions are mainly distributed in low-light and indoor scenes, while ours performed well in these challenging scenes.

(2) As the resolution decreases, the success rates of learning-based methods decrease while ours remains robust.

Besides, to perceive the robustness more intuitively, Fig. 11 demonstrates two challenging examples in the scenes of indoors and dark. Since the sample in dark is too dark to see clearly, we impose image augmentation to better exhibit these results (Row 3 in Fig. 11). These examples are challenging for the feature-based solutions because the features in these

TABLE II
COMPARISON OF ROBUSTNESS FOR IMAGE STITCHING. THE NUMBER OF TESTING CASES IS 1,106

Input resolution	Metrics	Feature-based			Learning-based (supervised)		Learning-based (unsupervised)
		Homography [10]	APAP [13]	robust ELA [18]	VFISNet [35]	EPISNet [36]	Ours
512×512	Error	0	3	0	-	0	0
	Failure	86	31	111	-	22	15
	Total	86	34	111	-	22	15
	Success rate	92.22%	96.93%	89.96%	-	**98.01%**	**98.64%**
256×256	Error	0	10	0	-	0	0
	Failure	88	40	124	-	22	15
	Total	88	50	124	-	22	15
	Success rate	92.04%	95.48%	88.79%	-	**98.01%**	**98.64%**
128×128	Error	1	158	9	0	0	0
	Failure	206	66	214	131	32	15
	Total	207	224	223	131	32	15
	Success rate	81.28%	79.75%	79.84%	88.16%	**97.11%**	**98.64%**

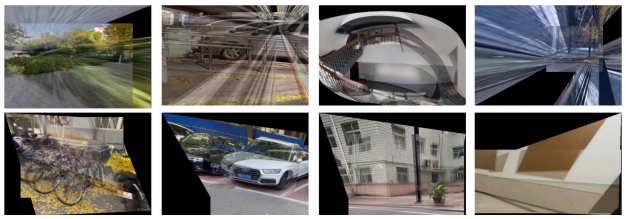


Fig. 10. Demonstration of 'failure'. Top: significant distortion. Bottom: intolerable artifacts.

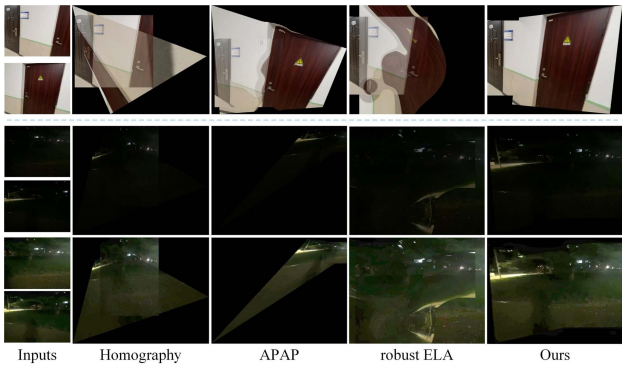


Fig. 11. Challenging samples to compare the robustness more intuitively in the scenes of indoors and dark. Row 1: indoors. Row 2: dark. Row 3: image augmentation to the dark scene. The resolution of the inputs is 512 × 512.

scenes are hard to detect. In contrast, our solution stitches them successfully due to the fantastic feature extraction capabilities of CNNs.

b) Study on Visual Quality: The proposed deep image stitching framework should be regarded as a whole which takes two images from arbitrary views as inputs and outputs the stitched result. Therefore, the traditional indicator that calculates the similarity of the overlapping regions is not suitable for our method. To compare with other methods quantitatively, we design user studies on visual quality. Specifically, we compare our method with Homography, APAP, and robust ELA one by one. At each time, four images are shown on one screen: the inputs, our stitched result, and the result from Homography/APAP/robust ELA. The results of ours and the other method are illustrated in random order each time. The user may zoom-in on the images and is required to answer which result is preferred. In the case of "no preference," the user needs to answer whether the two results are "both good" or "both bad". The studies are carried out in our testing set, which means every user has to compare each method with ours in 1,106 images. In this study, we invite 20 participants, including 10 researchers/students with computer vision backgrounds and 10 volunteers outside this community.

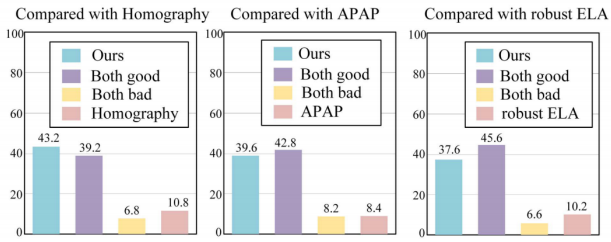


Fig. 12. User study on visual quality: compared with feature-based methods. The numbers are shown in percentage and averaged on 20 participants.

The results are shown in Fig. 12. Neglecting the ratio of both good and both bad, we find that preferring ours is significantly more than preferring other methods, which means our results have higher visual quality in users' evaluation.

To further demonstrate our performance, we also display the stitched results on the proposed real dataset (row 1-8 in Fig. 13) and on the classic image stitching instances outside of our dataset (row 9-10 in Fig. 13). All the cases are with varying degrees of parallax. Besides promising visual quality, it verifies the generalization ability of our model.

2) Compared with Learning-Based Solutions: The existing learning-based image stitching methods (VFISNet [35] and EPISNet [36]) are supervised learning methods, which require extra labels to train the network. In the case that it is unfair to compare our unsupervised solution with the supervised ones, our method still exhibits a superiority over them on robustness, continuity, illumination, and visual quality.

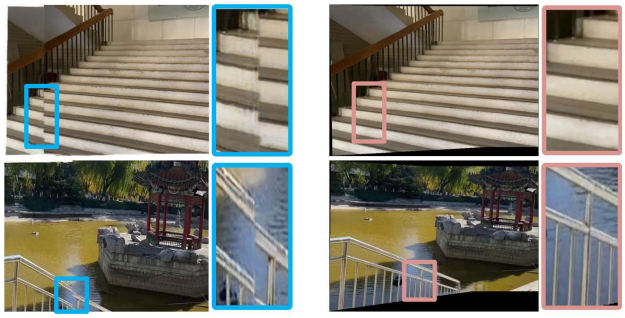
a) Study on Robustness: VFISNet is the first deep image stitching work that can stitch images from arbitrary views in a complete deep learning framework. However, it has a nonnegligible shortcoming: it can only stitch images of 128 × 128. Therefore, only the result under the resolution of 128 × 128 is given when measuring its robustness. The detailed results in Table II shows that the robustness of ours is better than other supervised ones. This can be accounted for by the following two reasons: (1) Our unsupervised deep homography model outperforms the other methods on robustness,

Fig. 13. Visual comparison of the image stitching quality. Row 1-8: instances with varying degrees of parallax from the proposed dataset. Row 9-10: "yard" [24] and "temple" [11] (classic image stitching instances outside of our dataset).

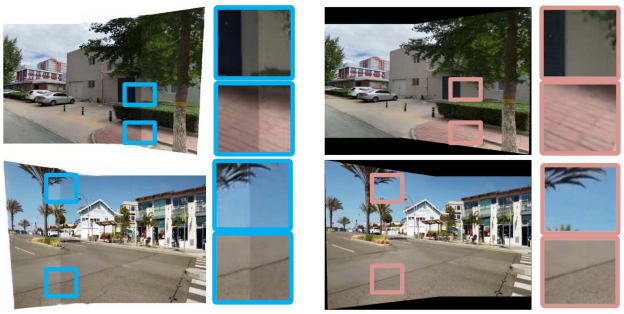
which significantly reduces failure cases caused by inaccurate homography estimation.

(2) Our unsupervised deep image reconstruction model can effectively reduce artifacts by reconstructing the stitched image from feature to pixel, which reduces failure cases caused by intolerant artifacts.

b) Study on Continuity: The supervised deep image stitching methods [35], [36] sacrifice the continuity of the edges (the edges between the reference image and the non-overlapping areas of the target image) to minimize artifacts. Although an edge-preserved network is proposed in EPISNet to weaken this problem, this problem still exists in a



(a) Comparison of edge continuity. Left: EPISNet [36]. Right: ours.



(b) Comparison of illumination difference. Left: EPISNet [36]. Right: ours.

Fig. 14. Study on continuity and illumination.

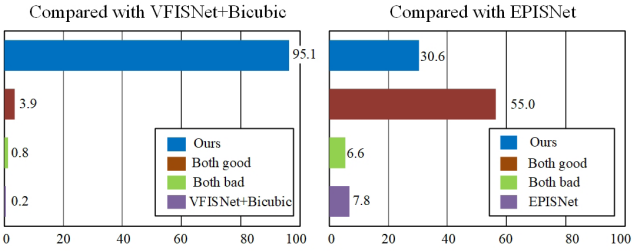


Fig. 15. User study on visual quality: compared with learning-based methods. The numbers are shown in percentage and averaged on 20 participants.

few testing cases. The discontinuity is demonstrated in the left picture of Fig. 14(a), where discontinuous areas are framed and enlarged. This problem is settled perfectly in our unsupervised approach, as shown in the right picture of Fig. 14(a). It gives credit to our constraint on seam masks, which enforces the edges of the overlapping areas close to one of the warped images.

c) Study on Illumination: Another advantage of our method is that ours can smooth the illumination difference between the two images. The comparison with EPISNet are illustrated in 14 (b). The supervised methods fail to smooth the illumination difference because they are trained in a synthetic dataset with no illumination difference in the input images (the supervised methods cannot be trained in a real dataset due to the lack of stitched labels). On the contrary, our method is trained in real scenes, which can effectively learn how to smooth the illumination difference caused by different shooting positions.

d) Study on Visual Quality: Similar to the user study with feature-based methods, we adopt the same strategy to investigate every participant to compare our method with

TABLE III
FRAMEWORKS FOR ABLATION STUDIES

	Architecture		Loss		
	LR branch	HR branch	Content loss	Seam loss	CS loss
v1	✓		✓		
v2	✓	✓	✓		
v3	✓	✓	✓	✓	
Ours	✓	✓	✓	✓	✓

the existing learning-based ones. Considering VFISNet can only work on the resolution of 128×128, we use Bicubic interpolation to resize the stitched images. The results are shown in Fig. 15. Since Bicubic interpolation inevitably brings blurs when zooming in on images, the probability of preferring our method is further greater than that of preferring VFISNet+Bicubic. Even compared with EPISNet, our method is still preferred on the visual quality of the stitched images.

Besides that, Fig. 13 exhibits the visual comparative results with these supervised methods, where the green rectangles indicate the severely blurred regions and the red rectangles point to discontinuous edges.

To perceive our visual quality more intuitively, more results are illustrated in Fig. 16, where the inputs and the outputs are demonstrated together.

D. Ablation Studies

In this section, ablation studies are performed on both network architectures and loss functions. In the architecture, we validate the effectiveness of the low-resolution branch (LR branch) and high-resolution branch (HR branch); in the loss, we test the function of the content loss, seam loss, and content consistency loss (CS loss). The properties of all the studied frameworks are shown in Table III.

From the results which are illustrated in Fig. 17, we can observe:

(1) The most straightforward combination of LR branch and content loss can realize image stitching. However, there are still two issues unresolved: seam distortions (row 1, col 4 in Fig. 17) and limited resolution. In our analysis, the seam distortion is the side effect of the proposed content loss.

(2) Compared v2 with v1, the HR branch can effectively enhance the resolution of the stitched image. As the cost, a few artifacts (row 2, col 2 in Fig. 17) are introduced since the receptive field of HR branch convolution kernels is too small for higher resolution images.

(3) Compared with v2, v3 removes the seam distortions (row 3, col 4 in Fig. 17) using the proposed seam loss. By imposing a pixel-level similarity constraint on the edge of the overlapping area, the seam distortions are suppressed successfully. However, there are still artifacts (row 3, col 2 in Fig. 17) in the stitched image.

(4) Compared with v3, ours removes the artifacts (row 4, col 2 in Fig. 17) using the proposed CS loss. The CS loss serves as an enhancer of the receptive field, which promotes the receptive field of the HR branch from that of the LR branch.

(a) Results on classic instances outside of our dataset. From left to right: "roof" [51], "theater" [20], "street" [52], "roadside" [14], and "officedesk" [20].

(b) Results on our proposed dataset. From left to right: "stairs", "snow", "grass", "lake", and "campus".

Fig. 16. More results of ours.

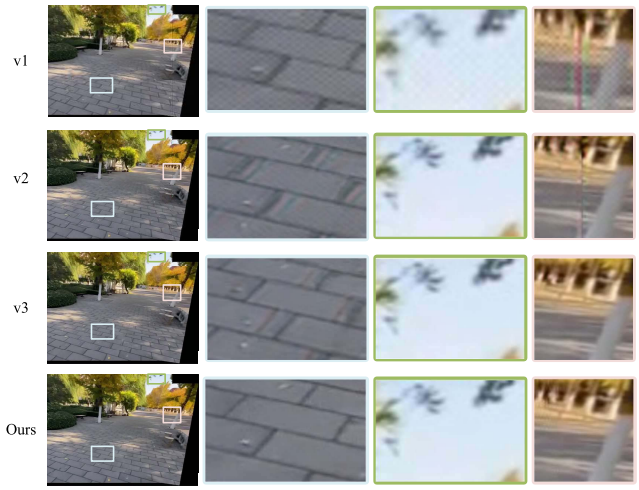


Fig. 17. Ablation studies on our framework. Col 1: outputs of different frameworks. Col 2-4: enlarged image patches to show the differences on artifacts, definition, and seam distortions, respectively.

VI. Limitation and Future Work

The proposed solution eliminates parallax artifacts through reconstructing the stitched images from feature to pixel. It is still essentially a stitching method based on a single homography. As the parallax increases, the alignment performance of the first stage will decrease, while the burden of the reconstruction network will also become heavier. When the parallax is too large, the reconstruction network may treat the misalignments as new objects to reconstruct. An example is shown in Fig. 18. In the future, we hope to solve this problem in two directions: 1) Improve the alignment performance of the alignment network to decrease the burden of the reconstruction network. 2) Increase the receptive field of the reconstruction network to deal with remained large misalignments.

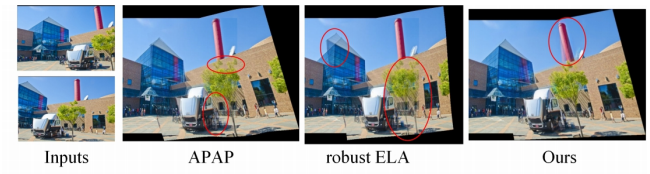


Fig. 18. A failure example. The red circle indicates the unsatisfying stitched areas.

VII. Conclusion

This paper proposes an unsupervised deep image stitching framework, comprising unsupervised coarse image alignment and unsupervised image reconstruction. In the alignment stage, an ablation-based loss function is proposed to constrain the unsupervised deep homography estimation in large-baseline scenes, and a stitching-domain transformer layer is designed to warp the input images in the stitching-domain space. In the reconstruction stage, an unsupervised deep image reconstruction network is proposed to reconstruct the stitched images from feature to pixel, eliminating the artifacts in an unsupervised reconstruction manner. Besides, a real dataset for unsupervised deep image stitching is presented, which we hope can work as a benchmark dataset for other methods. Experimental results demonstrate the superiority of our method over other state-of-the-art solutions. Even if compared with the supervised deep image stitching solutions, the results of our unsupervised approach are still preferred by users in terms of visual quality.

However, the reconstruction ability is not unlimited, which indicates our solution may fail in the scenes with extremely large parallax. Considering our first stage is essentially an alignment model based on a single homography, the ability to handle large parallax can be improved by extending the linear deep homography network to a non-linear homography model. Moreover, the reconstruction performance can be further increased by increasing the receptive field of the reconstruction network, which is also an exploring direction of the future work.

REFERENCES

[1] J. Chalfoun *et al.*, "MIST: Accurate and scalable microscopy image stitching tool with stage modeling and error minimization," *Sci. Rep.*, vol. 7, no. 1, pp. 1–10, Dec. 2017.

[2] E. Semenishchev, V. V. Voronin, V. I. Marchuk, and I. V. Tolstova, "Method for stitching microbial images using a neural network," *Proc. SPIE*, vol. 10221, May 2017, Art. no. 102210O.

[3] D. Li, Q. He, C. Liu, and H. Yu, "Medical image stitching using parallel SIFT detection and transformation fitting by particle swarm optimization," *J. Med. Imag. Health Informat.*, vol. 7, no. 6, pp. 1139–1148, Oct. 2017.

[4] J. Li, Y. Zhao, W. Ye, K. Yu, and S. Ge, "Attentive deep stitching and quality assessment for 360° omnidirectional images," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 1, pp. 209–221, Nov. 2019.

[5] V. R. Gaddam, M. Riegler, R. Eg, C. Griwodz, and P. Halvorsen, "Tiling in interactive panoramic video: Approaches and evaluation," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1819–1831, Sep. 2016.

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

[7] W.-S. Lai, O. Gallo, J. Gu, D. Sun, M.-H. Yang, and J. Kautz, "Video stitching for linear camera arrays," 2019, *arXiv:1907.13622*. [Online]. Available: https://arxiv.org/abs/1907.13622

[8] R. Anderson *et al.*, "Jump: Virtual reality video," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 1–13, Nov. 2016.

[9] H. G. Kim, H.-T. Lim, and Y. M. Ro, "Deep virtual reality image quality assessment with human perception guider for omnidirectional image," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 4, pp. 917–928, Apr. 2020.

[10] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[11] J. Gao, S. J. Kim, and M. S. Brown, "Constructing image panoramas using dual-homography warping," in *Proc. CVPR*, Jun. 2011, pp. 49–56.

[12] W.-Y. Lin, S. Liu, Y. Matsushita, T.-T. Ng, and L.-F. Cheong, "Smoothly varying affine stitching," in *Proc. CVPR*, Jun. 2011, pp. 345–352.

[13] J. Zaragoza, T.-J. Chin, M. S. Brown, and D. Suter, "As-projective-as-possible image stitching with moving DLT," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2339–2346.

[14] C.-H. Chang, Y. Sato, and Y.-Y. Chuang, "Shape-preserving half-projective warps for image stitching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3254–3261.

[15] C.-H. Chang and Y.-Y. Chuang, "A line-structure-preserving approach to image resizing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1075–1082.

[16] Y.-S. Chen and Y.-Y. Chuang, "Natural image stitching with the global similarity prior," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 186–201.

[17] C.-C. Lin, S. U. Pankanti, K. N. Ramamurthy, and A. Y. Aravkin, "Adaptive as-natural-as-possible image stitching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1155–1163.

[18] J. Li, Z. Wang, S. Lai, Y. Zhai, and M. Zhang, "Parallax-tolerant image stitching based on robust elastic warping," *IEEE Trans. Multimedia*, vol. 20, no. 7, pp. 1672–1687, Jul. 2018.

[19] K.-Y. Lee and J.-Y. Sim, "Warping residual based image stitching for large parallax," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 8198–8206.

[20] J. Li, B. Deng, R. Tang, Z. Wang, and Y. Yan, "Local-adaptive image alignment based on triangular facet approximation," *IEEE Trans. Image Process.*, vol. 29, pp. 2356–2369, Oct. 2020.

[21] A. Eden, M. Uyttendaele, and R. Szeliski, "Seamless image stitching of scenes with large motions and exposure differences," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 2498–2505.

[22] F. Zhang and F. Liu, "Parallax-tolerant image stitching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3262–3269.

[23] K. Lin, N. Jiang, L.-F. Cheong, M. Do, and J. Lu, "Seagull: Seam-guided local alignment for parallax-tolerant image stitching," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 370–385.

[24] J. Gao, Y. Li, T.-J. Chin, and M. S. Brown, "Seam-driven image stitching," in *Eurographics*. Wiley, 2013, pp. 45–48.

[25] H. Hejazifar and H. Khotanlou, "Fast and robust seam estimation to seamless image stitching," *Signal, Image Video Process.*, vol. 12, no. 5, pp. 885–893, 2018.

[26] A. Zomet, A. Levin, S. Peleg, and Y. Weiss, "Seamless image stitching by minimizing false edges," *IEEE Trans. Image Process.*, vol. 15, no. 4, pp. 969–977, Apr. 2006.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[28] Z. Zhang, C. Xu, J. Yang, J. Gao, and Z. Cui, "Progressive hard-mining network for monocular depth estimation," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3691–3702, Aug. 2018.

[29] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8934–8943.

[30] L. Tian, Z. Tu, D. Zhang, J. Liu, B. Li, and J. Yuan, "Unsupervised learning of optical flow with CNN-based non-local filtering," *IEEE Trans. Image Process.*, vol. 29, pp. 8429–8442, Aug. 2020.

[31] K. Liao, C. Lin, Y. Zhao, and M. Xu, "Model-free distortion rectification framework bridged by distortion distribution map," *IEEE Trans. Image Process.*, vol. 29, pp. 3707–3718, Jan. 2020.

[32] V.-D. Hoang, D.-P. Tran, N. G. Nhu, and V.-H. Pham, "Deep feature extraction for panoramic image stitching," in *Proc. Asian Conf. Intell. Inf. Database Syst.* Cham, Switzerland: Springer, 2020, pp. 141–151.

[33] 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, 2020.

[34] 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)*, Aug. 2019, pp. 192–197.

[35] 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.

[36] L. Nie, C. Lin, K. Liao, and Y. Zhao, "Learning edge-preserved image stitching from large-baseline deep homography," 2020, *arXiv:2012.06194*. [Online]. Available: https://arxiv.org/abs/2012.06194

[37] 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.

[38] J. Zhang *et al.*, "Content-aware unsupervised deep homography estimation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 653–669.

[39] A. Agarwala *et al.*, "Interactive digital photomontage," in *Proc. ACM SIGGRAPH Papers*, 2004, pp. 294–302.

[40] F. Liu, M. Gleicher, H. Jin, and A. Agarwala, "Content-preserving warps for 3D video stabilization," *ACM Trans. Graph.*, vol. 28, no. 3, pp. 1–9, 2009.

[41] F. Zhang and F. Liu, "Casual stereoscopic panorama stitching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2002–2010.

[42] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Deep image homography estimation," 2016, *arXiv:1606.03798*. [Online]. Available: https://arxiv.org/abs/1606.03798

[43] H. Le, F. Liu, S. Zhang, and A. Agarwala, "Deep homography estimation for dynamic scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7652–7661.

[44] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.

[45] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput.-Assisted Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[46] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 694–711.
[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980
[49] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[50] M. A. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.
[51] Y. Zhang, Y.-K. Lai, and F.-L. Zhang, "Content-preserving image stitching with piecewise rectangular boundary constraints," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 7, pp. 3198–3212, Jul. 2021.
[52] B. He and S. Yu, "Parallax-robust surveillance video stitching," *Sensors*, vol. 16, no. 1, p. 7, Dec. 2015.

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 (Member, IEEE) received the Ph.D. degree from Beijing Jiaotong University (BJTU), Beijing, China, in 2011.

From 2009 to 2010, he was a Visiting Researcher with the ICT Group, Delft University of Technology, The Netherlands. From 2011 to 2012, he was a Postdoctoral Researcher with the Multimedia Laboratory, Gent University, Belgium. He is currently a Professor with BJTU. His research interests include image/video compression and robust transmission, 3D video coding, virtual reality video processing, and ADAS.

Kang Liao received the B.S. degree in software engineering from Shaanxi Normal University, Xi'an, China, in 2017, 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 scene understanding, and adversarial learning.

Shuaicheng Liu (Member, IEEE) received the B.E. degree from Sichuan University, Chengdu, China, in 2008, and the M.S. and Ph.D. degrees from the National University of Singapore, Singapore, in 2010 and 2014, respectively. Since 2014, he has been an Associate Professor with the Institute of Image Processing, School of Information and Communication Engineering, University of Electronic Science and Technology of China (UESTC). His research interests include computer vision and computer graphics.

Yao Zhao (Senior Member, IEEE) received the B.S. degree from the Radio Engineering Department, Fuzhou University, Fuzhou, China, in 1989, the M.E. degree from the Radio Engineering Department, Southeast University, Nanjing, China, in 1992, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), Beijing, China, in 1996.

He became an Associate Professor with BJTU in 1998, where he 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. He was elected as a Chang Jiang Scholar of Ministry of Education of China in 2013. He is a Fellow of the IET.