Learning Pixel-wise Alignment for Unsupervised Image Stitching

Qi Jia
Dalian University of Technology
jiaqi@dlut.edu.cn

Xiaomei Feng
Dalian University of Technology
xiaomeifeng19@gmail.com

Yu Liu*
Dalian University of Technology
liuyu8824@dlut.edu.cn

Xin Fan
Dalian University of Technology
xin.fan@dlut.edu.cn

Longin Jan Latecki
Temple University
latecki@temple.edu

ABSTRACT

Image stitching aims to align a pair of images in the same view. Generating precise alignment with natural structures is challenging for image stitching, as there is no wider field-of-view image as a reference, especially in non-coplanar practical scenarios. In this paper, we propose an unsupervised image stitching framework, breaking through the coplanar constraints in homography estimation, yielding accurate pixel-wise alignment under limited overlapping regions. First, we generate a global transformation by an iterative dense feature matching combined with an error control strategy to alleviate the difference introduced by large parallax. Second, we propose a pixel-wise warping network embedded within a large-scale feature extractor and a correlative feature enhancement module to explicitly learn correspondences between the inputs, and generate accurate pixel-level offsets upon novel constraints on both overlapping and non-overlapping regions. Notably, we leverage the pixel-level offsets in the overlapping area to guide the adjustment in the non-overlapping area upon content and structure consistency constraints, rendering a natural transition between two regions and distortions suppression over the entire stitched image. The proposed method achieves state-of-the-art performance that surpasses both traditional and deep learning approaches by a large margin. It also achieves the shortest execution time and has the best generalization ability on the traditional dataset.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**;

KEYWORDS

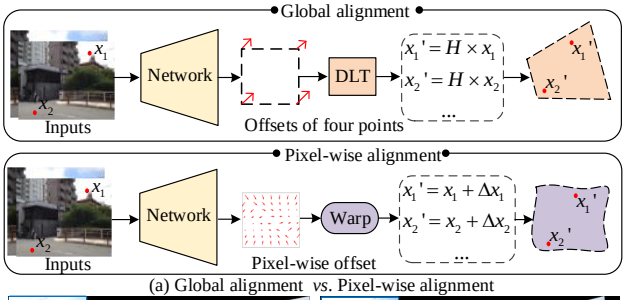
image stitching, pixel-wise alignment, homography estimation

ACM Reference Format:
Qi Jia, Xiaomei Feng, Yu Liu, Xin Fan, and Longin Jan Latecki. 2023. Learning Pixel-wise Alignment for Unsupervised Image Stitching. In *Proceedings of*

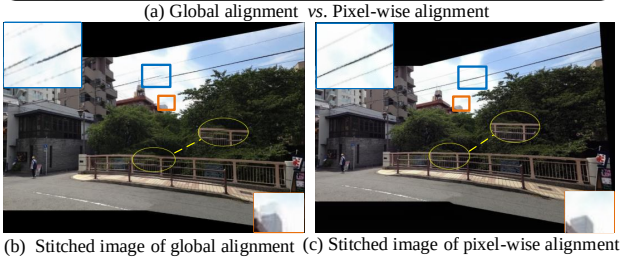
*Corresponding author: Yu Liu.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612298



(a) Global alignment *vs.* Pixel-wise alignment



(b) Stitched image of global alignment (c) Stitched image of pixel-wise alignment

Figure 1: Global alignment *vs.* Pixel-wise alignment. (a) Illustration of the difference between global and pixel-wise alignment in principle. Global alignment applies a direct linear transformation (DLT) to approximate the offset for all the pixels, while our pixel-wise alignment executes uniform transformations for each pixel. (b) and (c) compare two kinds of stitching results. Pixel-wise alignment achieves superior results in image and artifact suppression compared to global alignment, as shown in the zoomed-in regions.

the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3612298

1 INTRODUCTION

Image stitching aims to estimate an accurate transformation between a pair of images and align them in the same view. It has been a well-studied topic with widespread applications [38] such as panorama on smartphones [42], robot navigation [7], and virtual reality [1, 18]. However, generating high-quality stitched images in various practical scenarios is still challenging, especially when there is no wider field-of-view image as a reference.

Homography transformation [5, 9, 44] is the most widely used image stitching model, that leverages the feature correlation in overlapping regions as constraints to estimate a global homography matrix [30], and transform the whole target image to the view of the reference image (see the global warping part in Fig. 1 (a)). Most existing methods estimate the global homography by assuming the whole scene is coplanar, leading to severe misalignment

and artifacts in the stitching result [30], as shown in the zoomed-in region of Fig. 1 (b). However, in theory, homography transformation only holds for the coplanar scenario, whereas hardly exists in practice. Therefore, both conventional [3, 12, 20, 21, 25] and deep learning methods [6, 28, 45] are devoted to finding an approximate solution to obtain an accurate alignment. To reduce the impact of non-coplanarity on homography estimation, some works divide the image into multiple uniform patches as approximate coplanar regions to calculate multiple homography transformations, such as conventional dual-homography warp (DHW) [12], and multi-homography method as-projection-as-possible (APAP) [43]. However, there are at least three main limitations for existing homography estimation based methods: (1) the global or divided image patches have no coplanar guarantee, which is only an approximate solution; (2) the global alignment estimates a single or limited number of homography transformations for the whole image, which is insufficient to achieve pixel-wise accurate alignment, as shown in Fig. 1 (a); (3) there are no real stitched results as a reference, which is a challenge for training deep learning methods. Consequently, it is crucial to achieve a nonuniform alignment per pixel.

In contrast, existing pixel-level alignment methods are only applicable to almost complete overlapping image pairs, such as medical image registration [46] or registration between consecutive video frames [13]. They estimate the pixel-level offsets of the whole image in a low resolution by searching for feature correlation between image pairs from the whole image. Unlike image registration, image stitching has limited overlapping regions and large parallax, lacking constraints for non-overlapping regions in an unsupervised framework. Consequently, existing image registration methods are unable to produce correct offsets for non-overlapping regions, and fail to output the whole image stitching result. Moreover, image registration employs a global feature-matching strategy, which is prone to induce mismatched features in limited overlapping regions of image stitching.

In this paper, we propose a coarse-to-fine unsupervised image stitching network to achieve pixel-wise alignment. First, we estimate a global homography to handle large-scale viewpoint variations, rendering a uniform alignment for input image pairs. Second, we explore both texture and geometric consistency constraints to achieve non-uniform pixel-wise alignment in the overlapping area. Moreover, we leverage overlapping regions to guide the consistency of non-overlapping regions in content and structure to adjust the alignment for the whole stitched image. Our method demonstrates promising performance on stitched images, rendering fewer artifacts and misalignment as presented in the zoomed-in regions of Fig. 1 (c). Numerous qualitative and quantitative results validate the effectiveness of the proposed method. Our contributions are three-fold:

- We propose a coarse-to-fine unsupervised image stitching framework to align pixels starting from a uniform transformation and moving to anisotropic pixel-wise offsets, which breaks through the coplanar constraints of a single homography for the first time.
- We design an overlapping region-guided pixel-wise warping network with a large-scale feature extractor and a correlative feature enhancement module to capture pixel-level correspondences, invoking accurate alignment with high-resolution offsets.
- We leverage pixel-wise alignment of overlapping regions to guide the adjustment of non-overlapping regions in an unsupervised manner, preserving consistent structure and content for the whole stitched images under the condition of a limited extent of overlapping regions.

Our method outperforms both traditional and deep learning state-of-the-art approaches by a large margin and has the shortest execution time on all challenging datasets with visually superior stitching results. In particular, it has 34.42% lower alignment errors on average than that of the existing best method [21]. Sections 3 and 4 elaborate on our contributions.

2 RELATED WORK

Traditional image stitching methods. Traditional image stitching methods usually estimate an optimal global transformation by matching anchors. SIFT [26] and SURF [2] are widely used to detect and match feature points, followed by RANdom SAmple Consensus (RANSAC) [10] to estimate a homography for image pairs. As the single homography transformation only works well for ideal coplanar scenarios, some methods try to provide adaptive warping schemes for different non-coplanar regions [3, 12, 20, 25]. However, undesired distortion still occurs for large parallax images. To reduce the distortions and artifacts introduced by limited homography estimation, APAP [43] estimates homography for multiple patches to cover the warping of different regions. Subsequently, Liao *et al.* [24] propose single-perspective warping (SPW) that leverages both point and line pairs as anchors. Jia *et al.* [17] consider the local coplanar relations of line-point pairs (LPC) that leverage the coplanarity of matching line-point pairs to align images while suppressing distortion in non-overlapping regions. In addition, Du *et al.* [8] propose a geometric structure-preserving stitching method (GES-GSP). However, parameter setting has a serious impact on such traditional methods, making them sensitive to parallax changes. Particularly, traditional methods require high computation complexity to detect and matching features, while they fail easily when a limited number of matched features are present.

Deep learning-based image stitching methods. Deep learning methods are more adaptive than traditional methods on homography estimation, as the powerful representation learning of convolutional neural networks can yield dense matching features [14, 35], even in low-texture images. In addition, deep learning-based homography estimation yields favorable outcomes on synthetic images [6, 29, 31] or small parallax datasets [45]. Nevertheless, synthetic images still assume that the whole scenario is coplanar, which hardly exists in reality.

As there are no real stitched images as ground truth, some methods employ inner or outer constraints to reduce distortion and artifacts in real scenes with large parallax. The outer constraints are related to the fixed relative positions of cameras, which are widely used in autonomous driving [19, 40] and video surveillance [22]. The inner constraints refer to the feature correlation between image pairs. Based on this, an unsupervised image stitching framework (UDIS) is proposed [30] that employs DLT to produce the global homography, however, these existing methods still operate under the

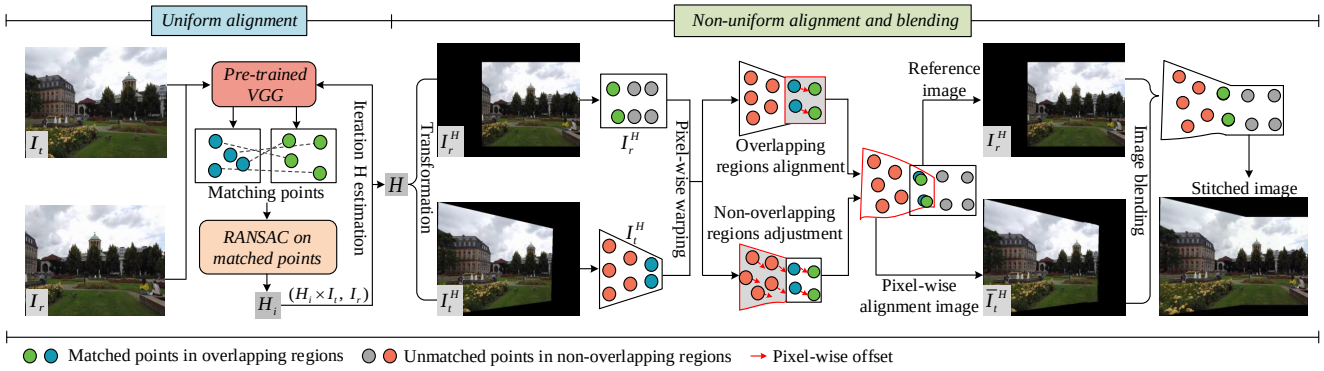


Figure 2: Pipeline of the proposed unsupervised pixel-wise image stitching framework. In the uniform pre-alignment phase, we warp the target image I_t to the view of the reference image I_r by a homography transformation H. Then, pre-aligned image pairs are fed into the non-uniform alignment network to generate pixel-wise alignment image $\bar{I}_t^H$. We use different constraints to align the overlapping regions and non-overlapping regions. Finally, we leverage a seamless image blending module to reduce the difference in brightness.

coplanar hypothesis, which leads to artifacts and distortions due to uniform alignment. *Unlike them, our proposed unsupervised image alignment framework solves the problem by estimating the pixel-wise offsets, rendering anisotropic accurate alignment for non-coplanar scenarios in reality.*

Image registration. Image registration is mainly applied to a non-rigid alignment for two images with large overlapping areas. The image deformation is considered as the motion of the object, and the displacement field of the registration is estimated by computing the instantaneous speed field of each pixel [15]. In [13, 34, 39, 46], optical flow is proposed to align the video adjacent frames in a supervised manner. However, the estimation of the dense displacement field is vulnerable to failure in the case of large parallax [37]. Consequently, it is impossible to apply a dense flow field to image stitching, due to a huge difference in the overlapping region, and the lack of constraints in non-overlapping regions. *In contrast to previous approaches, we leverage the alignment of the overlapping region to guide the alignment of non-overlapping regions, rendering consistent structure and texture for the entire stitched image.*

3 PIXEL-WISE IMAGE STITCHING

The proposed image stitching pipeline is illustrated in Fig. 2. Firstly, given a pair of target image I_t and reference image I_r, we conduct a uniform pre-alignment by estimating a global homography H in an iterative manner to reduce the impact of large parallax (Section 3.1). On top of the pre-aligned image I_r^H and I_t^H, we design a non-uniform alignment network that explores the correlation of image features and generates the pixel-wise alignment image $I_t^{\bar{H}}$ (Section 3.2). To achieve the structure and texture consistency of the whole stitching image, we propose a series of constraints for overlapping and non-overlapping regions by weighted masks (Section 3.3). Finally, we introduce spatial and channel attentions successively [11] on a U-Net structure as a blending module to adjust the brightness and color of images from different viewpoints.

3.1 Uniform Pre-alignment by Homography

We leverage the pre-trained VGG model [36] to extract and match dense features in different layers [9] and define an error evaluation index combined with RANSAC to iteratively perform feature matching and global homography estimation, as illustrated in the uniform alignment of Fig 2. To preserve the linear and textural structure of the original image, we employ a similarity transformation matrix to control the distortion during the homography transformation. We estimate the similarity matrix H_s by the four corresponding corner points of the target image before and after warping. We measure the distortion by computing how homography transformation H_i deviates from its best-fitting similarity transformation H_s. Let P^l be the four corner points of the target image, $l = 1, 2, 3, 4$. The error-index $\hat{E}_i$ is calculated by:

$$\hat{E}_i = \arg\min_{H_i} \sum_{l=1}^{4} \left\| H_s P^l - H_i P^l \right\|^2, \tag{1}$$

where H_i is the estimated homography matrix in the *i-th* iteration. When $\hat{E}_i$ is smaller than a threshold (0.01 in our paper), the iteration is terminated, which usually stops within three iterations for most image pairs. Pre-alignment is an efficient process to reduce the parallax of image pairs, such that the following pixel-wise warping can be more accurate.

3.2 Non-uniform Alignment for Pixel-wise Offset

Based on Section 3.1, the anisotropic alignment network is designed to refine the pixel-wise offsets of the whole image, as illustrated in Fig. 3. Our input images include the whole pre-aligned image I_r^H, I_t^H, and image of overlapping regions $I_t^H \cap I_r^H$. Firstly, we employ a feature extractor $\mathcal{F}$ to obtain the features of input images. Subsequently, a correlative feature enhancement module $C\mathcal{F}E$ is employed to increase the weight of overlapping regions, and the enhanced features are fed into the correlation computation module CC to compute the feature correlation. Finally, the pixel-wise offset estimation module outputs the final alignment results. Below, we elaborate the design of each module.

Feature extractor. To obtain high-resolution pixel-level offsets, we design three cascaded convolution layers to output large-scale feature maps. We employ residual blocks and 7×7 convolution kernel to increase the receptive field [34] in the convolution process.

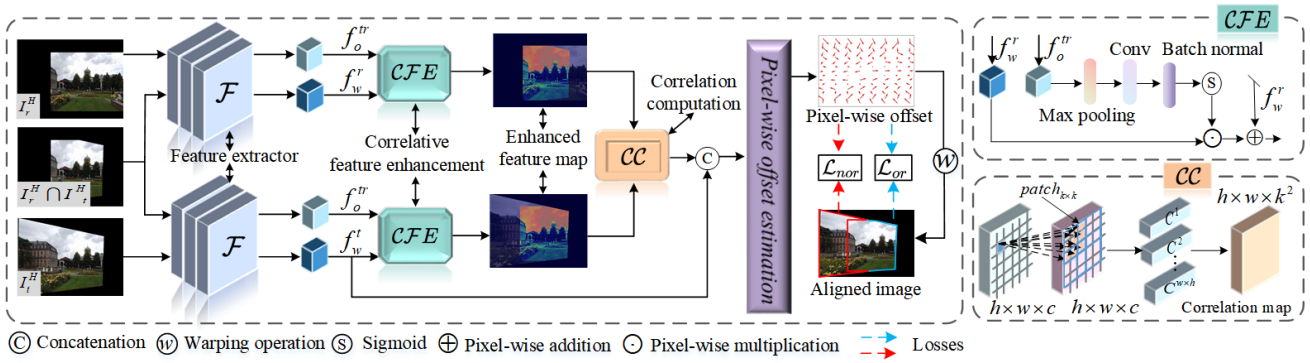


Figure 3: The architecture of non-uniform pixel-wise alignment network. The input of the network includes the pre-aligned reference image I_r^H, target image I_t^H, and overlapping regions $I_r^H \cap I_t^H$, where f_w^r, f_w^t, and f_{or}^{tr} indicate their feature maps extracted by feature extractor $\mathcal{F}$. The correlative features enhancement module $\mathcal{CFE}$ consolidates the features in overlapping regions to achieve accurate matching in the correlation computation module $\mathcal{CC}$. C^1, C^2, ... $C^{w\times h}$ represent the pixel similarity tensor in $\mathcal{CC}$. $\mathcal{L}_{or}$ and $\mathcal{L}_{nor}$ represent the constraints of overlapping regions and non-overlapping regions, respectively.

The size of extracted feature f_w^r, f_w^t, and f_o^{tr} are $H/2 \times W/2$, where H and W are the initial dimensions of the inputs.

Correlative feature enhancement ($\mathcal{CFE}$). This module aims to enhance the features in overlapping regions, as matching features only exist in overlapping regions and the matching accuracy directly affects the precision of offset estimation. The top right of Fig. 3 demonstrates the structure of $\mathcal{CFE}$. The process is attention-guided to increase interest in overlapping regions. Firstly, the overlapping feature f_o^{tr} is converted to probability maps via max pooling, convolution, batch normalization, and the sigmoid layer. Then, the probability map strengthens the overlapping regions of the whole image feature f_w^t by dot product. Finally, the enhanced features combine the original feature f_w^t by pixel-wise addition to restore non-overlapping features. Eq. 2 illustrates the whole process.

$$F^r = f_w^r \oplus (\mathcal{S}(f_o^{tr}) \odot f_w^r), \tag{2}$$

where $\mathcal{S}(\cdot)$ represents features after sigmoid layer, $\odot$ indicates pixel multiplication, and $\oplus$ is pixel-wise addition. f_o^{tr} and f_w^t also follow the same process to produce correlative enhanced features F^t.

Correlation computation ($\mathcal{CC}$). The enhanced features F^t and F^r are fed into the correlation computation module to capture the similarities of features, as illustrated at the bottom right corner of Fig. 3. To reduce the impact of mismatched features, we employ L^2-normalization on extracted features to yield better distinctions on matching features. Each pixel in F^r is compared to pixels in a $k \times k$, $k = 7$ square neighborhood in the F^t to generate similarity tensor C^i, $i \in w \times h$ by cosine similarity [33]. The cosine similarity s is defined as:

$$s < (m,n),(u,v) >= \frac{F^r(m,n) \cdot F^t(u,v)}{\|F^r(m,n)\| \, \|F^t(u,v)\|}, \tag{3}$$

where $F^r(m,n)$ and $F^t(u,v)$ represent the pixel on feature map of F^r and F^t at positions (m,n) and (u,v), respectively. In this way, we obtain a final $h \times w \times k^2$ similarity tensor for all pixels in F^r and F^t, and w, h represent the size of the feature map.

Pixel-wise offset estimation and blending. We concatenate the similarity tensor and the feature f_w^t of the target image as input to estimate the pixel-wise offset. We employ Conv+BN+ReLU convolutional blocks and upsampling operation to generate a dense displacement field $\mathbf{F}_{t\to r}$ of the same size as the input image. The pixel-level offset $\mathbf{F}_{t\to r}$ transforms I_t^H to produce pixel-wise alignment image $\bar{I}_t^H$.

To preserve the consistency of brightness and color in overlapping and non-overlapping regions, we employ a blending module to generate clear and natural stitching images. As image blending is related to both local and global features of the entire stitched image, we develop the U-Net structure by introducing spatial and channel attention mechanisms in the convolutional layers to integrate both local and global features. As image blending is outside our main contributions, we outline the module in our manuscript.

3.3 Loss Function

Our goal is to design constraints that align overlapping regions accurately, while smoothly transiting the overlapping regions to the non-overlapping regions. The total loss function for jointly training overlapping $\mathcal{L}_{or}$ and non-overlapping regions $\mathcal{L}_{nor}$ is designed as:

$$\mathcal{L}_{total} = \mathcal{L}_{or} + \mathcal{L}_{nor}. \tag{4}$$

Objective loss for overlapping regions. For ideal alignment, the overlapping regions are supposed to fit perfectly. The loss for overlapping regions $\mathcal{L}_{or}$ is composed of three constraints: a matching consistency $\mathcal{L}_{match}$, a texture consistency $\mathcal{L}_{ssim}$ and a geometric cyclic consistency $\mathcal{L}_{geoc}$, which is defined as:

$$\mathcal{L}_{or} = \lambda_{or}\mathcal{L}_{match} + \mathcal{L}_{ssim} + \mu_{or}\mathcal{L}_{geoc}, \tag{5}$$

where λ_{or} and μ_{or} are hyper-parameters, indicating the weights of different losses. Each loss term is defined at the pixel level and is described in detail below with the pixel v_i in I_t^H and the corresponding pixel v_i' in I_r^H.

Matching consistency constraint. To obtain robust matching features, we generate matching probability maps by performing sigmoid regression on the correlation computation block. $\mathcal{M}_{t\to r}^{(v_i)}$ is the predicted matching probability from $I_t^H(v_i)$ to $I_r^H(v_i')$, and $\mathcal{M}_{r\to t}^{(v_i')}$ is the reverse. Ideally, the matching probability of $\mathcal{M}_{t\to r}^{(v_i)}$ and $\mathcal{M}_{r\to t}^{(v_i')}$ are consistent for matching pixel pairs. Therefore, we encourage this cycle-consistent matching to be close to 1, and the $\mathcal{L}_{match}$ is defined as:

$$\mathcal{L}_{match} = \sum_{v_i \in I_t^H} \left| \mathcal{M}_{t\to r}^{v_i} \odot \mathcal{M}_{r\to t}^{v_i'} - 1 \right|, \tag{6}$$

where $\odot$ indicates pixel-wise multiplication. As pixels in the overlapping regions have different contributions to the alignment, we apply $\mathcal{M}_{cyc} = M_{t\to r} \odot M_{r\to t}$ as the pixel-level weights for both $\mathcal{L}_{ssim}$ and $\mathcal{L}_{geoc}$.

Texture consistency constraint. During warping, the target image is supposed to be close to the pre-aligned reference image I_r^H. Therefore, we introduce the structural similarity (SSIM) constraint [34, 41] to comprehensively evaluate the texture similarity of the overlapping regions *w.r.t* the brightness, contrast, and structure. We define the loss as:

$$\mathcal{L}_{ssim} = \sum_{v_i \in I_t^H} \mathcal{M}_{cyc}^{v_i} (1 - SSIM(\mathbf{F}_{t\to r}^{v_i} \textcircled{w} I_t^H(v_i), I_r^H(v_i'))), \tag{7}$$

where $\textcircled{w}$ indicates warping operation, $\mathbf{F}_{t\to r}$ represents the estimated pixel offset from I_t^H to I_r^H. We sum over all pixels and take the average as the loss.

Geometric cyclic consistency constraint. To restrain the dense displacement between I_r^H and I_t^H, we can also get the offset field $\mathbf{F}_{r\to t}$ from reference to the target image by exchanging their position in the input. Then, we use $\mathbf{F}_{r\to t}$ to warp pixel v_i' in I_r^H to target the image viewpoint. Ideally, $\mathbf{F}_{r\to t}^{v_i'} \textcircled{w} v_i'$ is supposed to coincide with v_i in I_t^H if the estimated pixel offset is accurate, with the formula below:

$$\mathcal{L}_{geoc} = \sum_{v_i \in I_t^H} \mathcal{M}_{cyc}^{v_i} \odot \left\| v_i, \mathbf{F}_{r\to t}^{v_i'} \textcircled{w} v_i' \right\|_2. \tag{8}$$

Objective loss for non-overlapping regions. Overlapping regions have inherent constraints between image pairs, while there is no reference for the non-overlapping regions. To this end, we propose to leverage the offset field of overlapping regions to guide non-overlapping regions for smooth transiting. In addition, we employ content consistency to preserve the naturalness of the non-overlapping regions during warping. The total loss is defined as:

$$\mathcal{L}_{nor} = \mathcal{L}_{str} + \mathcal{L}_{con}, \tag{9}$$

where $\mathcal{L}_{str}$ and $\mathcal{L}_{con}$ represent structure and content constraints, respectively.

Structural consistency constraint. We design structural consistency constraints to control the pixel offset in non-overlapping regions. As the pixel-by-pixel offset is within a certain range, the offset in the same plane should be relatively smooth. As illustrated in Fig. 4, the offsets for non-overlapping regions are disordered compared to the overlapping region, which leads to a local unnatural stretching or compression of the warped image. Therefore, we encourage neighboring blocks to have similar motion trends to suppress distortions. Since the maximum pixel offset of our network is within $d = 24$ pixels, we divide the pixel offset field into W/d blocks, and the loss is defined as:

$$\mathcal{L}_{str} = \frac{1}{T} \sum_{i=1}^{W/d} \left\| \vec{b}_{i-1} + \vec{b}_{i+1} - 2\vec{b}_i \right\|_2, \tag{10}$$

where $\vec{b}_{i-1}$, $\vec{b}_i$, $\vec{b}_{i+1}$ represent average offsets in three successive blocks respectively, and $T = W/d - 2$ is the total number of computed successive blocks offset tuples.

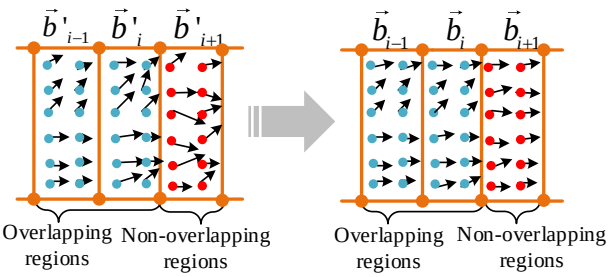


Figure 4: Diagram of the structural consistency constraint, w/o $\mathcal{L}_{str}$ on the left and w/ $\mathcal{L}_{str}$ on the right. The arrows indicate the directions of pixel movement, and the length indicates the size of the displacement. The offsets converge gradually from the left figure to the right one due to the structural consistency constraint. $\vec{b}_{i-1}$, $\vec{b}_i$, $\vec{b}_{i+1}$ represent average offsets in three consecutive blocks along x-axis.

Content consistency constraint. We also use content consistency constraints to suppress the distortion that occurs in non-overlapping regions during warping. As all the pixels in the non-overlapping region follow the same protocol, we use a binary mask $\mathcal{M}_{nor}$ to select non-overlapping regions features. Based on the output of pre-alignment, $\mathcal{M}_{nor} = \mathcal{M}_t \oplus (\mathcal{M}_r \cap \mathcal{M}_t)$, where $\mathcal{M}_r$ and $\mathcal{M}_t$ are transformed image regions, and $\oplus$ means exclusive OR. We use L_2 norm to constrain the content:

$$\mathcal{L}_{con} = \left\| I_t^H \odot \mathcal{M}_{nor}, \bar{I}_t^H \odot \mathcal{M}_{nor} \right\|_2, \tag{11}$$

where $\bar{I}_t^H$ represents the pixel-wise warped target image, and I_t^H represents the pre-aligned target image.

4 EXPERIMENTS

4.1 Implementation details

Datasets. We trained our network on a natural image dataset (UDIS-D) [30], which contains 10440 training images and 1106 test images with a size of 512×512. In addition, we also collect a challenging traditional dataset (Tra dataset) with a severe change of views as a cross-dataset to validate the generalization ability of our method. The Tra dataset consists of 100 images collected from SUA [25] , APAP [43], SPHP [3] , DHW [12] , DFW [23], REW [20], GSP [4]. In general, the stitching of the Tra dataset is more challenging as they have a larger parallax than the UDIS-D dataset, and we have no training on this dataset.

Details. We resize the pre-aligned images to 224×224 for the pixel-wise warping network during the training phase, but there is no size limit for our test images. We train our pixel-wise warping networks for 100 epochs using Adam optimizer [27] with the learning rate of 2×10^{-4}. We set $\lambda_{or} = 0.01$ and $\mu_{or} = 1$ for the constrain of overlapping regions. The whole framework is implemented in PyTorch framework [32] with an NVIDIA RTX 3090 GPU.

Evaluation metrics. We employ RMSE [43], PSNR [16], and SSIM [41] to evaluate the performance and follow the same evaluation protocol with other image stitching methods [17, 30]. RMSE indicates the matching error between the matched points. PSNR and SSIM are utilized to measure the feature similarity between the warped target and reference images in the overlapping region. We compare the overlapping regions of the stitching domain instead of

Table 1: Comparison on the UDIS-D dataset. The results show the average performance on the test set measured with PSNR, SSIM, and RMSE. The first and second-best solutions are marked in red and in blue, respectively.

Methods	PSNR ↑				SSIM ↑				RMSE ↓			
	Easy	Moderate	Hard	Average	Easy	Moderate	Hard	Average	Easy	Moderate	Hard	Average
APAP [43]	28.31	24.38	20.66	23.77	0.950	0.923	0.793	0.866	0.891	1.488	4.455	2.669
SPW [24]	26.61	23.03	20.00	23.07	0.932	0.867	0.760	0.849	1.209	1.986	4.478	2.645
ELA [21]	**29.38**	**25.91**	**21.53**	**24.84**	**0.960**	**0.933**	**0.853**	**0.904**	**0.574**	**0.928**	**2.383**	**1.496**
LPC [17]	27.84	23.75	20.45	23.39	0.945	0.890	0.780	0.855	1.166	1.725	4.753	2.934
CA-UDHN [45]	18.05	13.62	11.11	14.22	0.719	0.523	0.354	0.530	4.347	15.564	33.542	17.998
UDIS [30]	27.22	23.13	19.93	22.78	0.935	0.860	0.720	0.816	2.529	3.238	6.148	4.377
Ours	**31.31**	**27.18**	**23.49**	**26.61**	**0.972**	**0.950**	**0.906**	**0.936**	**0.458**	**0.630**	**1.250**	**0.868**

the original domain for all methods as our pixel-wise alignment is working on the whole stitching domain image. Besides, we resize all the warped images to the same size to compare for fairness.

4.2 Comparison with Existing Methods

Compared methods. We compare the proposed method with two categories of existing stitching methods: 1) Traditional methods including the classical method APAP [43], and three recent methods SPW [24], LPC [17], and ELA [21]. 2) Unsupervised deep learning-based methods, including CA-UDHN [45], UDIS [30]. We perform qualitative and quantitative comparisons with these methods on the testing set from UDIS-D and Tra datasets, covering diverse types of camera motion, scene, and field of view.

Table 2: The average PSNR, SSIM, and RMSE of our method to SOTA methods on the Tra dataset. The first and second-best solutions are marked in red and blue, respectively.

Methods	APAP	SPW	ELA	LPC	UDIS	Ours
PSNR ↑	**25.85**	24.39	25.39	24.70	21.73	**26.68**
SSIM ↑	**0.931**	0.904	0.924	0.910	0.848	**0.949**
RMSE ↓	2.570	4.357	**1.961**	4.531	37.218	**1.286**

Quantitative comparison. We report the performance of our method and the six most related state-of-the-art (SOTA) methods on the UDIS-D dataset in Table 1, where the first four rows are the classical multi-homography method (APAP [43]) and mesh optimization methods (SPW [24], ELA [21], LPC [17]), and the last two rows are deep learning methods. We divide the testing results into three levels, including 'Easy' (Top 0-30%), 'Moderate' (Top 30-60%), and 'Hard' (Top 60-100%). The average performance for each method is demonstrated in the last column of each metric. Table 1 demonstrates that our method outperforms all the other methods by a large margin on all metrics. Our method achieves the best RMSE of 0.868, which is 41.98% lower than that of the second-best method ELA, indicating the advantage of our pixel-level alignment over multi-homography estimation. Our method also achieves the best average PSNR of 26.61 dB, which is 1.77 dB higher than ELA. For the deep learning-based methods, CA-UDHN [45] is a global homography estimation method, which is only applicable to small parallax images, so it has the lowest values on the feature-based evaluation metrics PSNR and SSIM. Our method outperforms UDIS [30] by 16.81%, 14.71%, and 80.17% on average on PSNR, SSIM, and RMSE, respectively. It indicates the superiority of our pixel-wise warping in image alignment.

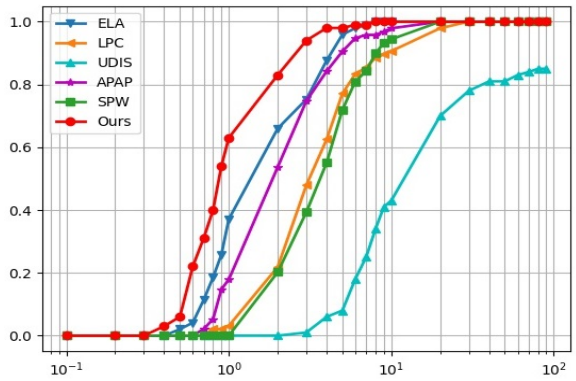


Figure 5: The matching error RMSE of different methods on the Tra dataset. The horizontal axis indicates the average matching error and the vertical axis indicates the percentage of images with RMSE value less than the value of the horizontal axis.

Validation on Cross-dataset. To test the generalization ability of the proposed method, we test our trained model on the collected cross-dataset of the Tra dataset, which exhibits a larger parallax than the UDIS-D dataset. Table 2 demonstrates our method achieves SOTA performance on all metrics. Our performance on RMSE is 1.286, reducing 34.42% compared with the second-best method ELA.

In addition, we demonstrate the overall RMSE distribution of different methods on the Tra dataset in Fig. 5. Take our performance marked by red points as an example, the value 10^0 on the x-coordinate with a corresponding value beyond 0.6 on the y-coordinate indicates more than 60% image pairs in the test set exhibit RMSE value less than 1, while the second best ELA method has only about 40%. The curve of our method distributes at the left-most with a big margin over other methods, indicating our method holds the absolute predominance on the cross dataset.

Qualitative comparison. In Fig. 6, we demonstrate qualitative comparisons of different methods on four image pairs, in which the first two rows are test images on the UDIS-D dataset and the last two rows from the Tra dataset. Our method surpasses all the other methods in the alignment quality with visually artifacts-free stitched images. The first two rows show challenging complex-texture and low-texture image pairs, respectively. The zoomed-in areas in the bottom show that the traditional methods APAP [43], SPW [24], ELA [21], and LPC [17] exhibit severe misalignment

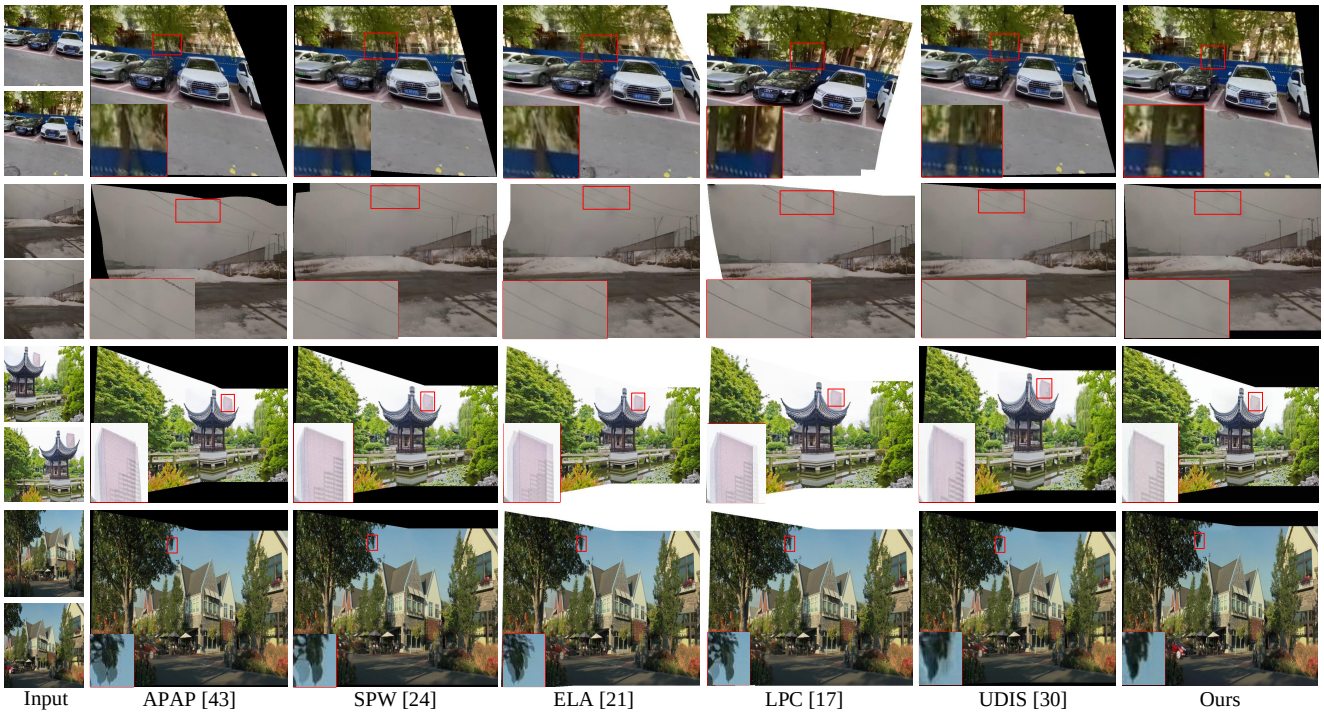


Figure 6: Comparison of different stitching methods on the UDIS-D dataset (rows 1 and 2) and the Tra dataset (rows 3 and 4), where the zoomed-in results are located below the stitched images and marked with red squares.

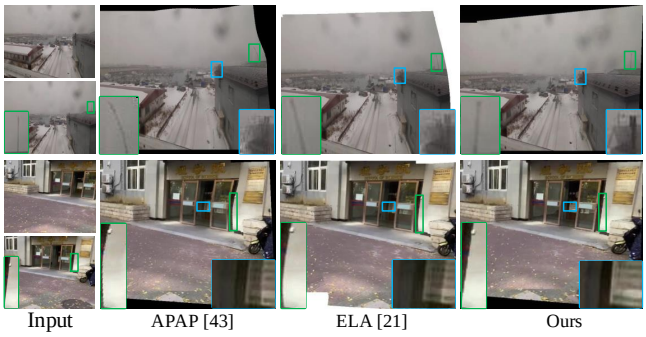


Figure 7: Comparison of the top three stitching methods on the UDIS-D dataset, where the zoomed-in regions in overlapping and non-overlapping regions are marked in blue and green squares, respectively.

in the tree trunk and antenna, due to their limitation in feature detection and matching. The deep learning method UDIS exhibits better performance than traditional methods but still suffers from artifacts as demonstrated in the zoomed-in areas. Image pairs in the last two rows have a large parallax, and objects in the overlapping regions exhibit different degrees of artifacts due to misalignment. The clear and natural stitching results of the proposed method demonstrate the benefits of our pixel-level warping, which enables us to handle the alignment of large parallax images with high accuracy.

Figure 7 compares the distortion of non-overlapping regions among the top three methods, including APAP and ELA. Different from the single homography transformation, APAP and ELA also adjust the non-overlapping regions. The green box marks non-overlapping regions and the blue box marks the overlapping regions. The zoomed-in regions at the bottom corners demonstrate our method surpasses the other two methods with the overlapping region being visually artifacts-free. At the same time, the inherent texture and structure in the non-overlapping regions are preserved. In contrast, the other two methods suffer from severe distortions, such as bent antenna and deformed walls in the green box. Consequently, our method not only aligns overlapping regions accurately but also suppresses the distortion of non-overlapping regions, indicating the effectiveness of the proposed content consistency and structural consistency constraints.

Table 3: Comparison of elapsed time(s) for different resolutions. The bold type indicates the best performance.

Resolution	1000 × 750	975 × 583	640 × 480	600 × 400
APAP [43]	19.417	6.408	20.291	5.534
SPW [24]	4.479	11.813	6.228	2.773
ELA [21]	4.067	3.192	1.938	2.106
LPC [17]	3.230	47.265	12.748	8.397
UDIS [30]	4.035	3.992	3.671	3.642
Ours	**2.689**	**2.653**	**1.370**	**1.168**

Computational efficiency comparison. We randomly select four images in the Tra dataset with different image resolutions. Table 3 demonstrates the average time of running each method five times under the same hardware configurations. Our elapsed time includes both pre-alignment and pixel-wise alignment stages, rendering the least execution time over all instances. Generally, the time cost reduces as the resolution decreases for most methods. There are exceptions when the number of matching features is independent

Figure 8: The comparison of coarse uniform pre-alignment and fine pixel-wise alignment. The green and blue squares indicate zoomed-in regions captured in overlapping regions.

of the image resolution, which mainly affects the computation times of traditional methods. Our average run time is only 49.36% of UDIS [30] over all resolutions, as our model size is only 18.5M compared with 2.1G of UDIS. In addition, the pre-alignment stage also cost very little time, as we use a pre-trained model for feature extraction.

Table 4: Ablation studies of different scale features on the UDIS-D dataset.

Variants	PSNR ↑	SSIM ↑	RMSE ↓
1/8 *scale*	24.80	0.886	2.669
1/4 *scale*	25.81	0.915	1.151
1/2 *scale*	**26.32**	**0.921**	**0.987**

Table 5: Ablation studies of correlative feature enhancement ($\mathcal{CFE}$) module on the UDIS-D dataset.

Variants	PSNR ↑	SSIM ↑	RMSE ↓
$\mathcal{S}(f_o^{tr}) \odot f_w^r$	26.22	0.927	1.026
$f_o^{tr} \oplus f_w^r$	26.56	0.928	0.921
$\mathcal{S}(f_o^{tr}) \odot f_w^r \oplus f_w^r$	**26.61**	**0.936**	**0.868**

4.3 Ablation Studies and Analysis

Performance of two-stage alignment. We conduct two alignment stages in our coarse-to-fine framework: uniform alignment in the pre-alignment stage and the non-uniform pixel-wise alignment stage. Fig. 8 illustrates the difference between the pre-alignment and our pixel-wise alignment for the image pair on the Tra dataset. Pre-alignment suffers from severe misalignment and artifacts, as shown in the zoomed-in green and blue boxes, respectively. In contrast, we obtain precisely aligned images when adding the second stage of non-uniform alignment, validating the necessity of pixel-wise alignment in non-coplanar scenes.

Effectiveness of large-scale features. To validate the effectiveness of large-scale feature maps, we change the feature extractor to extract different scale feature maps including 1/8 *scale*, 1/4 *scale*, and 1/2 *scale* of the input image. Table 4 demonstrates that the PSNR and SSIM increase when using larger sizes of feature maps. We obtain 26.32 dB on PSNR with $H/2 \times W/2$ size feature maps, which is significantly higher than 24.80 dB with $H/8 \times W/8$ size. It clearly shows that large-scale feature maps are beneficial for our pixel-wise alignment.

Effectiveness of correlative feature enhancement module ($\mathcal{CFE}$). To explore the best feature fusion manner for overlapping feature f_o^{tr} and the entire image features f_w^r and f_w^t, we use attention-guided fusion ($\mathcal{S}(f_o^{tr}) \odot f_w^r$) and direct addition ($f_w^r \oplus f_o^{tr}$) for comparison. As shown in Table 5, the attention-guided fusion ($\mathcal{S}(f_o^{tr}) \odot f_w^r$) with 26.22 dB is inferior to $f_w^r \oplus f_o^{tr}$ with 26.56 dB due to it neglecting the non-overlapping features. Therefore, we use the combined $\mathcal{S}(f_o^{tr}) \odot f_w^r \oplus f_w^r$ fusion to fully consider both regions and achieve the best performance on all metrics compared with the other variants.

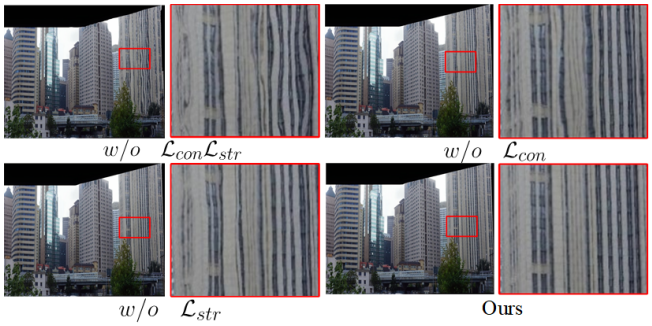


Figure 9: The visualization of different constraints on distortion suppression for the non-overlapping regions, which are marked with red boxes and shown in zoomed-in areas.

Effectiveness of constraints for non-overlapping regions. To validate the effectiveness of our constraints on content $\mathcal{L}_{con}$ and structure $\mathcal{L}_{str}$ for non-overlapping regions, Fig. 9 compares four image stitching results with and without corresponding constraints by the zoomed-in patches on the right. The upper left figure shows that the building texture and structure in the non-overlapping region are severely distorted without both losses. The quality of the following two stitching images significantly improves when adding either loss. The final result with both constraints exhibits the best visual quality with appealing well-organized structure and texture, rendering the effectiveness of the proposed loss terms, which do not require any supervision.

5 CONCLUSION

The proposed approach eliminates the coplanar limitations of a single homography by pixel-wise alignment. The proposed pixel-wise image stitching network employs a large-scale feature extractor and attention-guided modules to enhance the impact of the overlapping region to obtain high-resolution and accurate pixel-level offsets. A series of constraints is proposed for overlapping and non-overlapping regions to enforce consistency of features, contents, and structures between image pairs and in the stitched image. These novel constraints coupled with a specially designed network enable us to achieve precise alignment in an unsupervised manner. In comparison experiments, our model achieves the SOTA performance on the UDIS-D dataset and exhibits the best generalization ability on the Tra dataset compared with existing methods. In our future work, we will try to integrate global pre-alignment and pixel-wise stitching in an end-to-end network.

ACKNOWLEDGEMENTS

This work was supported in part by the NSF of China under Grant Nos. 62272083, 61876030 and 62102061, in part by the Liaoning Provincial NSF under Grant 2022-MS-128 and 2022-MS-137, and in part by the U.S. NSF under Grant IIS-2107213.

REFERENCES

[1] Robert Anderson, David Gallup, Jonathan T Barron, Janne Kontkanen, Noah Snavely, Carlos Hernández, Sameer Agarwal, and Steven M Seitz. 2016. Jump: virtual reality video. *ACM Transactions on Graphics (TOG)* 35, 6 (2016), 1–13.

[2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. 2006. Surf: Speeded up robust features. In *European conference on computer vision*. Springer, 404–417.

[3] Che-Han Chang, Yoichi Sato, and Yung-Yu Chuang. 2014. Shape-preserving half-projective warps for image stitching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3254–3261.

[4] Yu-Sheng Chen and Yung-Yu Chuang. 2016. Natural image stitching with the global similarity prior. In *European conference on computer vision*. Springer, 186–201.

[5] Xin Deng, Wenzhe Yang, Ren Yang, Mai Xu, Enpeng Liu, Qianhan Feng, and Radu Timofte. 2021. Deep homography for efficient stereo image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1492–1501.

[6] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2016. Deep image homography estimation. *arXiv preprint arXiv:1606.03798* (2016).

[7] Abhishek Kumar Dewangan, Rohit Raja, and Reetika Singh. 2014. An implementation of multi sensor based mobile robot with image stitching application. *Int J Comput Sci Mobile Comput* 3, 6 (2014), 603–609.

[8] Peng Du, Jifeng Ning, Jiguang Cui, Shaoli Huang, Xinchao Wang, and Jiaxin Wang. 2022. Geometric Structure Preserving Warp for Natural Image Stitching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3688–3696.

[9] Ufuk Efe, Kutalmis Gokalp Ince, and Aydin Alatan. 2021. Dfm: A performance baseline for deep feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4284–4293.

[10] Martin A Fischler and Robert C Bolles. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* 24, 6 (1981), 381–395.

[11] Xueyang Fu, Qi Qi, Zheng-Jun Zha, Yurui Zhu, and Xinghao Ding. 2021. Rain streak removal via dual graph convolutional network. In *Proc. AAAI Conf. Artif. Intell.* 1–9.

[12] Junhong Gao, Seon Joo Kim, and Michael S Brown. 2011. Constructing image panoramas using dual-homography warping. In *CVPR 2011*. IEEE, 49–56.

[13] Jun Han, Jun Tao, and Chaoli Wang. 2018. FlowNet: A deep learning framework for clustering and selection of streamlines and stream surfaces. *IEEE transactions on visualization and computer graphics* 26, 4 (2018), 1732–1744.

[14] Van-Dung Hoang, Diem-Phuc Tran, Nguyen Gia Nhu, Van-Huy Pham, et al. 2020. Deep feature extraction for panoramic image stitching. In *Asian Conference on Intelligent Information and Database Systems*. Springer, 141–151.

[15] Berthold KP Horn and Brian G Schunck. 1981. Determining optical flow. *Artificial intelligence* 17, 1-3 (1981), 185–203.

[16] Quan Huynh-Thu and Mohammed Ghanbari. 2008. Scope of validity of PSNR in image/video quality assessment. *Electronics letters* 44, 13 (2008), 800–801.

[17] Qi Jia, ZhengJun Li, Xin Fan, Haotian Zhao, Shiyu Teng, Xinchen Ye, and Longin Jan Latecki. 2021. Leveraging line-point consistence to preserve structures for wide parallax image stitching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12186–12195.

[18] Hak Gu Kim, Heoun-Taek Lim, and Yong Man Ro. 2019. Deep virtual reality image quality assessment with human perception guider for omnidirectional image. *IEEE Transactions on Circuits and Systems for Video Technology* 30, 4 (2019), 917–928.

[19] Wei-Sheng Lai, Orazio Gallo, Jinwei Gu, Deqing Sun, Ming-Hsuan Yang, and Jan Kautz. 2019. Video stitching for linear camera arrays. *arXiv preprint arXiv:1907.13622* (2019).

[20] Jing Li, Zhengming Wang, Shiming Lai, Yongping Zhai, and Maojun Zhang. 2017. Parallax-tolerant image stitching based on robust elastic warping. *IEEE Transactions on multimedia* 20, 7 (2017), 1672–1687.

[21] Jing Li, Zhengming Wang, Shiming Lai, Yongping Zhai, and Maojun Zhang. 2017. Parallax-tolerant image stitching based on robust elastic warping. *IEEE Transactions on multimedia* 20, 7 (2017), 1672–1687.

[22] Jia Li, Yifan Zhao, Weihua Ye, Kaiwen Yu, and Shiming Ge. 2019. Attentive Deep Stitching and Quality Assessment for 360 Omnidirectional Images. *IEEE Journal of Selected Topics in Signal Processing* 14, 1 (2019), 209–221.

[23] Shiwei Li, Lu Yuan, Jian Sun, and Long Quan. 2015. Dual-feature warping-based motion model estimation. In *Proceedings of the IEEE International Conference on Computer Vision*. 4283–4291.

[24] Tianli Liao and Nan Li. 2019. Single-perspective warps in natural image stitching. *IEEE transactions on image processing* 29 (2019), 724–735.

[25] Wen-Yan Lin, Siying Liu, Yasuyuki Matsushita, Tian-Tsong Ng, and Loong-Fah Cheong. 2011. Smoothly varying affine stitching. In *CVPR 2011*. IEEE, 345–352.

[26] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60, 2 (2004), 91–110.

[27] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. 2017. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2794–2802.

[28] Ty Nguyen, Steven W Chen, Shreyas S Shivakumar, Camillo Jose Taylor, and Vijay Kumar. 2018. Unsupervised deep homography: A fast and robust homography estimation model. *IEEE Robotics and Automation Letters* 3, 3 (2018), 2346–2353.

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

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

[31] Lang Nie, Chunyu Lin, Kang Liao, and Yao Zhao. 2020. Learning edge-preserved image stitching from large-baseline deep homography. *arXiv preprint arXiv:2012.06194* (2020).

[32] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. (2017).

[33] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. 2017. Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6148–6157.

[34] Xi Shen, François Darmon, Alexei A Efros, and Mathieu Aubry. 2020. Ransac-flow: generic two-stage image alignment. In *European Conference on Computer Vision*. Springer, 618–637.

[35] Zaifeng Shi, Hui Li, Qingjie Cao, Huizheng Ren, and Boyu Fan. 2020. An image mosaic method based on convolutional neural network semantic features extraction. *Journal of Signal Processing Systems* 92, 4 (2020), 435–444.

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

[37] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. 2018. PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[38] Richard Szeliski. 2006. Image alignment and stitching. In *Handbook of mathematical models in computer vision*. Springer, 273–292.

[39] Zachary Teed and Jia Deng. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*. Springer, 402–419.

[40] Lang Wang, Wen Yu, and Bao Li. 2020. Multi-scenes image stitching based on autonomous driving. In *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, Vol. 1. IEEE, 694–698.

[41] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[42] Yingen Xiong and Kari Pulli. 2009. Sequential image stitching for mobile panoramas. In *2009 7th International Conference on Information, Communications and Signal Processing (ICICS)*. IEEE, 1–5.

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

[44] Haoxian Zhang and Yonggen Ling. 2022. HVC-Net: Unifying Homography, Visibility, and Confidence Learning for Planar Object Tracking. In *European Conference on Computer Vision*. Springer, 701–718.

[45] Jirong Zhang, Chuan Wang, Shuaicheng Liu, Lanpeng Jia, Nianjin Ye, Jue Wang, Ji Zhou, and Jian Sun. 2020. Content-aware unsupervised deep homography estimation. In *European Conference on Computer Vision*. Springer, 653–669.

[46] Shengyu Zhao, Yue Dong, Eric I Chang, Yan Xu, et al. 2019. Recursive cascaded networks for unsupervised medical image registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10600–10610.