

# Deep Rectangling for Image Stitching: A Learning Baseline

Lang Nie<sup>1,2</sup>, Chunyu Lin<sup>1,2\*</sup>, Kang Liao<sup>1,2</sup>, Shuaicheng Liu<sup>3</sup>, Yao Zhao<sup>1,2</sup>

<sup>1</sup>Institute of Information Science, Beijing Jiaotong University, Beijing, China

<sup>2</sup>Beijing Key Laboratory of Advanced Information Science and Network, Beijing, China

<sup>3</sup>University of Electronic Science and Technology of China, Chengdu, China

<https://github.com/nie-lang/DeepRectangling>

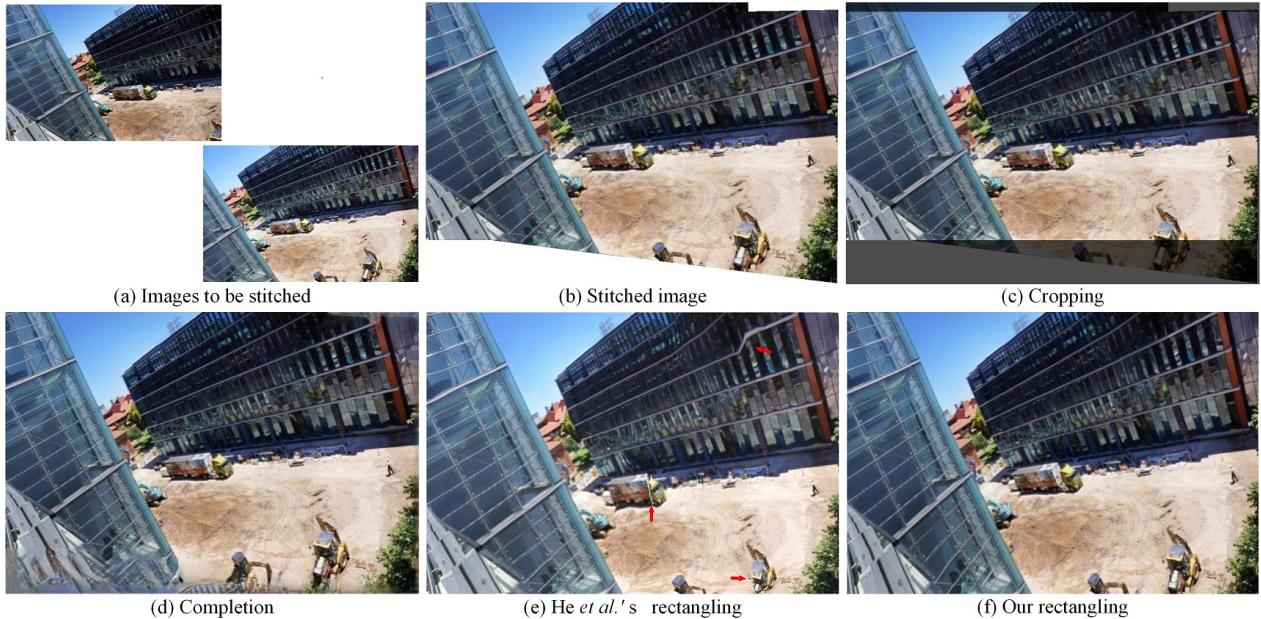


Figure 1. Different solutions to irregular boundaries in image stitching. (a) A classic image stitching dataset that is not included in the proposed dataset (APAP-conssite [30]). (b) Stitching images using UDIS [24] with inevitable irregular boundaries. (c) Cropping the boundaries to get a rectangular image. (d) Completing the missing regions using LaMa [26]. (e) He et al.’s rectangling [7] with noticeable distortions. (f) Our rectangling with high content fidelity.

## Abstract

Stitched images provide a wide field-of-view (FoV) but suffer from unpleasant irregular boundaries. To deal with this problem, existing image rectangling methods devote to searching an initial mesh and optimizing a target mesh to form the mesh deformation in two stages. Then rectangular images can be generated by warping stitched images. However, these solutions only work for images with rich linear structures, leading to noticeable distortions for portraits and landscapes with non-linear objects.

In this paper, we address these issues by proposing the first deep learning solution to image rectangling. Concretely, we predefine a rigid target mesh and only estimate an initial mesh to form the mesh deformation, contribut-

ing to a compact one-stage solution. The initial mesh is predicted using a fully convolutional network with a residual progressive regression strategy. To obtain results with high content fidelity, a comprehensive objective function is proposed to simultaneously encourage the boundary rectangling, mesh shape-preserving, and content perceptually natural. Besides, we build the first image stitching rectangling dataset with a large diversity in irregular boundaries and scenes. Experiments demonstrate our superiority over traditional methods both quantitatively and qualitatively.

## 1. Introduction

Image stitching algorithm [3, 13, 21, 30] can generate a wide FoV image (Fig.1b) from normal FoV images

(Fig.1a). These methods optimize a global or local warp to align the overlapping regions of different images. Nevertheless, non-overlapping regions always suffer from irregular boundaries [2]. People who use image stitching technology have to be tolerant of unpleasant boundaries.

To deal with the irregular boundaries, one of the solutions is to crop a stitched image with a rectangle. However, cropping inevitably reduces the FoV of the stitched image, which contradicts the original intention of image stitching. Fig.1c demonstrates an example, where the dark regions indicate the discarded areas by cropping. On the other hand, image completion can synthesize the missing regions to form a rectangular image. Nevertheless, there is currently no work to design a mask for irregular boundaries in image stitching, and even SOTA completion works [26, 28] show unsatisfying performance (Fig.1d) when processing the stitched images. Moreover, the completion methods may add some contents that seem to be harmonious but different from reality, making them unreliable in high-security applications such as autonomous driving [12].

To obtain a rectangular image with high content fidelity, image rectangling methods [6, 7, 14] are proposed to warp a stitched image to a rectangle via mesh deformation. However, these solutions can only preserve structures with straight/geodesic lines such as buildings, boxes, pillars, etc. For non-linear structures such as portraits [27], distortions are usually generated. Actually, the capability to preserve linear structures is limited by line detection, thus distortions also occur in linear structures sometimes (Fig.1e). Moreover, these traditional methods are two-stage solutions that search an initial mesh and optimize a target mesh successively, making it challenging to be parallelly accelerated.

To address the above problems, we propose the first one-stage learning baseline, in which we predefine a rigid target mesh and only predict an initial mesh. Specifically, we design a simple but effective fully convolutional network to estimate a content-aware initial mesh from a stitched image with a residual progressive regression strategy. Besides, a comprehensive objective function consisting of a boundary term, a mesh term, and a content term is proposed to simultaneously encourage the boundary rectangular, mesh shape-preserving, and content perceptually natural. Compared with the existing methods, our content-preserving capability is more general (not limited to linear structures) and more robust (Fig.1f) due to the effective semantic perception in our content constraint.

As there is no proper dataset readily available, we build a deep image rectangling dataset (DIR-D) to supervise our training. First, we apply He *et al.*'s rectangling [7] to real stitched images to generate synthetic rectangular images. Then we utilize the inverse of rectangling transformations to warp real rectangular images to synthetic stitched images. Finally, we manually filter out images without distortions

from tens of thousands synthetic images for several epochs strictly, yielding a dataset with 6,358 samples.

Experimental results show that our approach can generate content-preserving rectangular images efficiently and effectively, outperforming the existing solutions both quantitatively and qualitatively. To sum up, we conclude our contributions as follows:

- We propose the first deep rectangling solution for image stitching, which can effectively generate rectangular images in a residual progressive manner.
- Existing methods are two-stage solutions while ours is a one-stage solution, enabling efficient parallel computation with a predefined rigid target mesh. Besides, ours can preserve both linear and non-linear structures.
- As there is no proper dataset of pairs of stitched images and rectangular images, we build a deep image rectangling dataset with a wide range of irregular boundaries and scenes.

## 2. Related Work

This paper offers a deep learning based rectangling solution for image stitching. Hence, this section reviews previous works related to image stitching and image rectangling.

### 2.1. Image Stitching

Aligning overlapping regions [15] is the core goal of image stitching. But to produce natural stitched images, it is also necessary to minimize projective distortions of non-overlapping regions. In [2, 19], the projective transformation of overlapping regions is smoothly extrapolated into non-overlapping regions, and the resultant warp gradually changes from projection to similarity across the image. Li *et al.* [17] propose a quasi-homography warp, which relies on a global homography while squeezing non-overlapping areas. Liao and Li [18] propose two single-perspective warps to preserve perspective consistency with reduced projective distortions. Recently, Jia *et al.* [10] consider the scenes of long lines and keep the shape of global co-linear line segments during the stitching process.

Although the existing image stitching algorithms can reduce projective distortions and keep the natural appearance, they can not solve the problem of irregular boundaries in stitched images.

### 2.2. Image Rectangling

To get rectangular stitched images, He *et al.* [7] propose to optimize a line-preserving mesh deformation. However, the proposed energy function can only preserve linear structures. Considering straight lines may be bent in a panorama (ERP format), Li *et al.* [14] improve the line-preserving energy term into the geodesic-preserving energy term. But

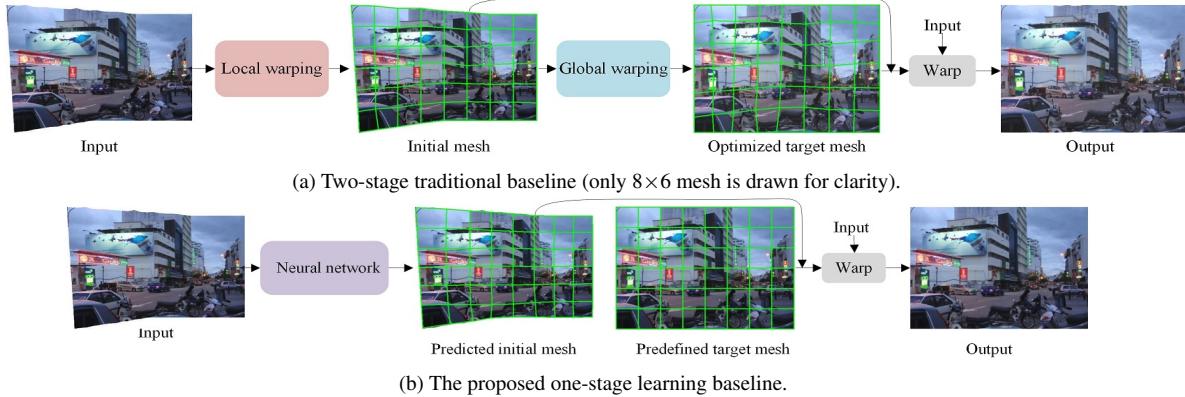


Figure 2. Traditional baseline vs. learning baseline. Traditional baseline solves the rectangling warp in two stages by searching the initial mesh and optimizing the target mesh successively, while our solution solves it in one stage because ours only predicts the initial mesh.

this improvement limits its application to the panorama and the geodesic lines can not be detected from a stitched image directly. Later, Zhang *et al.* [32] bridge the image rectangling and image stitching in a unified optimization. Nevertheless, to reduce distortions of the final rectangular result, they make a compromise to the rectangular shape, adopting piecewise rectangular boundary constraints instead.

Image rectangling is rarely studied because the unstable performance and heavy time-consumption make it impractical in applications. In this paper, we propose a simple but effective learning baseline to address these issues.

### 3. Methodology

We first analyse differences between the traditional baseline and the proposed learning baseline in Section 3.1. Then, our network structure and the objective function are discussed in Section 3.2 and Section 3.3, respectively.

#### 3.1. Traditional Baseline vs. Learning Baseline

A rectangling solution should solve the initial mesh and the target mesh to form the mesh deformation. Then the rectangling result can be obtained via warping.

##### 3.1.1 Traditional Baseline

In classic traditional methods [7, 14], two stages are required: local stage and global stage (shown in Fig. 2a).

**Stage 1: local stage.** First, insert abundant seams into the stitched image to get a preliminary rectangular image using seam carving algorithm [1]. Then, place a regular mesh on the preliminary rectangular image and remove all the seams to get an initial mesh for a stitched image with irregular boundaries.

**Stage 2: global stage.** This stage solves the optimal target mesh via optimizing an energy function to preserve limited perceptual properties such as straight lines.

They produce the rectangular image by warping the stitched image from the initial mesh to the target mesh.

##### 3.1.2 Learning Baseline

As shown in Fig. 2b, the proposed learning baseline is a one-stage solution.

Given a stitched image, our solution only needs to predict a content-aware initial mesh via a neural network. As for the target mesh, we predefined it to have a rigid shape. Moreover, the rigid mesh shape can enable the acceleration of the backward interpolation using the matrix computation easily [23]. Rectangular images can be obtained by warping stitched images from the predicted initial mesh to the predefined target mesh.

Compared with the traditional baseline, the learning baseline is more efficient due to the one-stage pipeline. The content-preserving capability makes our rectangling results more natural in perception (explained in Section 3.3.1).

#### 3.2. Network Structure

Similar to image completion tasks [26, 28], a stitched mask is also included in the input of the proposed network. As illustrated in Fig. 3, we concatenate the stitched image  $I$  and mask  $M$  on the channel dimension as the input. The output is the predicted mesh motions.

**Feature extractor.** We stack simple convolution-pooling blocks to extract high-level semantic features from the input. Formally, 8 convolutional layers are adopted, whose filter numbers are set to 64, 64, 64, 64, 128, 128, 128, and 128, respectively. The max-pooling layers are used after the 2-th, 4-th, and 6-th convolutional layers.

**Mesh motion regressor.** After feature extraction, an adaptive pooling layer is utilized to fix the resolution of feature maps. Subsequently, we design a fully convolutional structure as the mesh motion regressor to predict the horizontal and vertical motions of every vertex based on the regular mesh. Supposing the mesh resolution is  $U \times V$ , the size of the output volume is  $(U + 1) \times (V + 1) \times 2$ .

**Residual progressive regression.** Observing that the

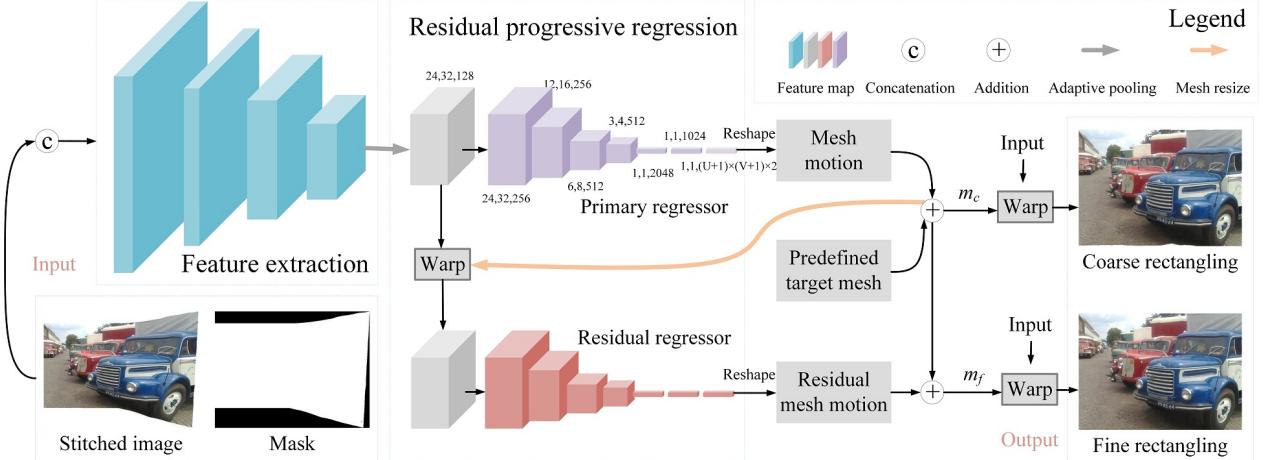


Figure 3. The overall structure of our network which takes a stitched image and a mask as input. It outputs the predicted mesh motions in a progressive manner. The rectangling results can be produced by warping the stitched image with the predicted warps.

warped result can be regarded as the network input again, we design a residual progressive regression strategy to estimate accurate mesh motions through a progressive manner. First, we do not use the warped image as the input of a new network directly, because this would double the computational complexity. On the contrary, we warp the intermediate feature maps instead, improving the performance with a slight increase in the computation. Then, we design two regressors with the same structure to predict primary mesh motions and residual mesh motions, respectively. Although they share the same structure, they are specified for different tasks due to the different input features.

### 3.3. Objective Function

We optimize our network parameters using a comprehensive objective function that consists of three terms. The optimization goal can be formulated as follows:

$$L_{total} = \ell_b + \ell_m + \ell_c, \quad (1)$$

where  $\ell_b$ ,  $\ell_m$ , and  $\ell_c$  are the boundary term, mesh term, and content term, respectively.

#### 3.3.1 Content Term

The traditional methods [7, 14] preserve image contents by preserving the angles of straight/geodesic lines, failing to deal with other non-linear structures. To overcome it, we propose to learn the content-preserving capability from two different perspectives.

**Appearance loss.** Given the predicted primary mesh  $m_p$  and final mesh  $m_f$ , we enforce the rectangling results to be close to the rectangling labels  $R$  in appearance as follows:

$$\ell_c^a = \|R - \mathcal{W}(I, m_p)\|_1 + \|R - \mathcal{W}(I, m_f)\|_1, \quad (2)$$

where  $\mathcal{W}(\cdot, \cdot)$  is the warp operation.

**Perception loss.** To make our results perceptually natural, we minimize the  $L2$  distance between rectangling results and labels in high-level semantic perception as Eq. 3:

$$\ell_c^p = \|\varphi(R) - \varphi(\mathcal{W}(I, m_c))\|_2 + \|\varphi(R) - \varphi(\mathcal{W}(I, m_f))\|_2, \quad (3)$$

where  $\varphi(\cdot)$  represent the operation of feature extraction from the 'conv4\_2' layer of VGG19 [25]. In this manner, various perceptual properties (not limited to linear structures) can be perceived.

In sum, the content loss is formed by simultaneously emphasizing the similarity in appearance and semantic perception as follows:

$$\ell_c = \omega_a \ell_c^a + \omega_p \ell_c^p, \quad (4)$$

where  $\omega_a$  and  $\omega_p$  denote the weights for the appearance loss and the perception loss.

#### 3.3.2 Mesh Term

To prevent content distortions in rectangular images, The predicted mesh should not be exaggeratedly deformed. Therefore, we design an intra-grid constraint and an inter-grid constraint to keep the shape of the deformed mesh.

**Intra-grid constraint.** In a grid, we impose constraints on the magnitude and direction of grid edges. As shown in Fig. 4a, we encourage the direction of the horizontal projection of each horizontal edge  $\vec{e}_u$  to the right, together with its norm greater than a threshold  $\alpha \frac{W}{V}$  (suppose the stitched image has the resolution of  $H \times W$ ). We use a penalty  $P_{hor}$  to describe this constraint as follows:

$$P_{hor} = \begin{cases} \alpha \frac{W}{V} - \langle \vec{e}_u, \vec{i} \rangle, & \langle \vec{e}_u, \vec{i} \rangle < \alpha \frac{W}{V} \\ 0, & \langle \vec{e}_u, \vec{i} \rangle \geq \alpha \frac{W}{V} \end{cases} \quad (5)$$

where  $i$  is the horizontal unit vector to the right. As for the vertical edge  $\vec{e}_v$  in every grid, we impose a similar penalty

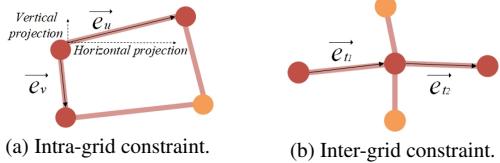


Figure 4. Mesh shape-preserving constraint.

$P_{ver}$  as follows:

$$P_{ver} = \begin{cases} \alpha \frac{H}{U} - \langle \vec{e}_v, \vec{j} \rangle, & \langle \vec{e}_u, \vec{j} \rangle < \alpha \frac{H}{U} \\ 0, & \langle \vec{e}_v, \vec{j} \rangle \geq \alpha \frac{H}{U} \end{cases} \quad (6)$$

where  $j$  is the vertical unit vector to the bottom. Then, the intra-grid mesh loss is formed using Eq. 7, which can effectively prevent the intra-grid shape from distortions.

$$\ell_m^{intra} = \frac{1}{(U+1) \times V} \sum_{\vec{e}_u \in m_p \cup m_f} P_{hor} + \frac{1}{U \times (V+1)} \sum_{\vec{e}_v \in m_p \cup m_f} P_{ver}. \quad (7)$$

**Inter-grid constraint.** We also adopt the inter-grid constraint to encourage neighboring grids to transform consistently. As shown in Fig. 4b, two successive deformed grid edges  $\{\vec{e}_{t1}, \vec{e}_{t2}\}$  are encouraged to be co-linear.

$$\ell_m^{inter} = \frac{1}{N} \sum_{\{\vec{e}_{t1}, \vec{e}_{t2}\} \in m_p \cup m_f} \left(1 - \frac{\langle \vec{e}_{t1}, \vec{e}_{t2} \rangle}{\|\vec{e}_{t1}\| \cdot \|\vec{e}_{t2}\|}\right). \quad (8)$$

We formulate the inter-grid mesh loss as above, where  $N$  is the number of tuples of two successive edges in a mesh.

In sum, the total mesh term is concluded as follows:

$$\ell_m = \ell_m^{intra} + \ell_m^{inter}, \quad (9)$$

### 3.3.3 Boundary Term

As for the boundary term, we constrain the mask instead of the predicted mesh. Given a 0-1 mask of a stitched image (as shown in Fig. 3), we warp the mask and constrain the warped mask close to an all-one matrix  $E$  as follows:

$$\ell_b = \|E - \mathcal{W}(M, m_p)\|_1 + \|E - \mathcal{W}(M, m_f)\|_1. \quad (10)$$

## 4. Data Preparation

To train a deep image rectangling network, we build an image rectangling dataset (DIR-D), in which each sample is a triplet consisting of a stitched image ( $I$ ), a mask ( $M$ ), and a rectangling label ( $R$ ). We prepare this dataset by the following steps:

**Step 1:** Adopt ELA [16] to stitch images from the UDIS-D dataset [24] to collect extensive real stitched images. Then we dismiss those with extrapolated areas less than 10% of the whole images.

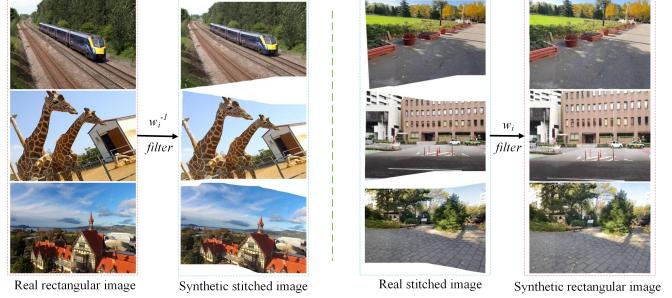


Figure 5. Dataset preparation. Left: synthesize stitched images from real rectangular images. Right: synthesize rectangular images from real stitched images. The synthesized samples will then undergo strict manual filtering to form a reliable dataset without noticeable distortions.

**Step 2:** Generate abundant different mesh deformation ( $w_i$ ) using He *et al.*'s algorithm [7] from these real stitched images as shown in Fig. 5 (right).

**Step 3:** Apply the inverse of the mesh deformation ( $w_i^{-1}$ ) to warp real rectangular images (from MS-COCO [20] and collected video frames) to synthetic stitched images as shown in Fig. 5 (left). The masks can be obtained by warping the all-one matrixes. Then we get triplets of real rectangular images ( $R$ ), synthetic stitched images ( $I$ ), and warped matrixes ( $M$ ).

**Step 4:** Dismiss the triplets whose  $I$  have distortions manually. Each manual operation will take 5-20s. Formally, we repeat this process for three epochs and 5,705 triplets remain from more than 60,000 samples.

**Step 5:** Mix real stitched images to our training set to increase the generalization capability. Specifically, we filter out 653 samples whose  $R$  has no distortion from more than 5,000 samples in step 2.

In sum, we prepare the DIR-D dataset with a wide range of irregular boundaries and scenes, which includes 5,839 samples for training and 519 samples for testing. Every image in the dataset has a resolution of  $512 \times 384$ .

## 5. Experiments

We first discuss experimental configuration and speed in Section 5.1. Then we demonstrate the comparative results and ablation studies in Section 5.2 and Section 5.3.

### 5.1. Experimental Configuration and Speed

Our network is trained using an Adam optimizer [11] with an exponentially decaying learning rate initialized to  $10^{-4}$  for  $100k$  iterations. The batch size is set to 4 and we use RELU as the activation function except that the last layers of regressors adopt no activation function.  $\omega_a, \omega_p$  and  $\alpha$  are assigned as 1,  $5e^{-6}$ , and 0.125, respectively.  $U \times V$  is set to  $8 \times 6$  and the implementation is based on TensorFlow. We use a single GPU with NVIDIA RTX 2080 Ti to finish

Table 1. Quantitative comparisons on DIR-D.

Method	FID [9] ↓	SSIM ↑	PSNR ↑
Reference	44.47	0.3245	11.30
He <i>et al.</i> 's. [7]	38.19	0.3775	14.70
Ours	<b>21.77</b>	<b>0.7141</b>	<b>21.28</b>

Table 2. No-reference blind image quality comparisons on DIR-D.

Method	BIQUE [29] ↓	NIQE [22] ↓
He <i>et al.</i> 's [7]	14.234	17.150
Ours	13.989	16.754
Label	<b>11.086</b>	<b>14.872</b>

all the training and inference.

It takes less than 0.4 seconds to process a 10 mega-pixel image. Similar to the experimental configuration of [7], the input image would be downsampled to 1 mega-pixel first and the mesh deformation is solved in the downsampled resolution. Then the mesh deformation would be upsampled and the rectangling result can be obtained by warping the full resolution input image using this upsampled deformation. The running time is dominantly on warping (interpolating) the full resolution image.

## 5.2. Comparative Result

To display our superiority comprehensively, we conduct comparative experiments in quantitative comparison, qualitative comparison, user study, and cross-dataset evaluation.

### 5.2.1 Quantitative Comparison

We compare our solution with He *et al.*'s method [7] on DIR-D, where 519 samples are tested for each method. We calculate the average FID [9], SSIM, and PSNR with labels to evaluate these solutions. The quantitative results are shown in Table 1, where ‘Reference’ takes stitched images as rectangling results for reference.

From this table, the proposed learning solution is significantly better than the traditional solution in every metric on DIR-D. This remarkable improvement is attributed to our content-preserving property that can preserve both linear and non-linear structures. Besides, when the location of an object in a rectangular result changes a little, it looks natural as well but the metrics might differ, which makes the quantitative experiments not completely convincing. Therefore, we further conduct a comparison of blind image quality evaluation. As shown in Table 2, we adopt BIQUE [29] and NIQE [22] as ‘no-reference’ assessment metrics, where our solution generate higher quality results. These evaluation methods are opinion-unaware methodologies that attempt to quantify the distortion without the need for any training data. We add the evaluation of ‘Label’ for reference, which indicates the upper limit of the performance.

### 5.2.2 Qualitative Comparison

To compare the qualitative results comprehensively, we divide the testing set into two parts. The first part includes

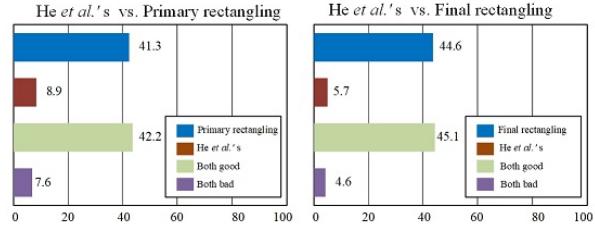


Figure 6. User study on DIR-D (519 testing samples). The numbers are shown in percentage and averaged on 10 participants.

abundant linear structures that are suited to the traditional baseline, while the second one includes extensive non-linear structures such as portraits.

From the results in Fig. 7, we can observe that our method significantly outperforms the traditional solution in the two scenes. We owe our superiority to the content preserving capability that can keep the mesh shape-preserving and content perceptually natural. The traditional solution does not perform well in scenes with linear structures due to the limited line detection capability. The failure occurs in portraits with non-linear objects because non-linear properties are not included in its optimized energy.

### 5.2.3 User Study

The motivation of image rectangling is that the users are not satisfied with the irregular boundaries in stitched images. Therefore, our goal is to produce rectangular images that please the most users.

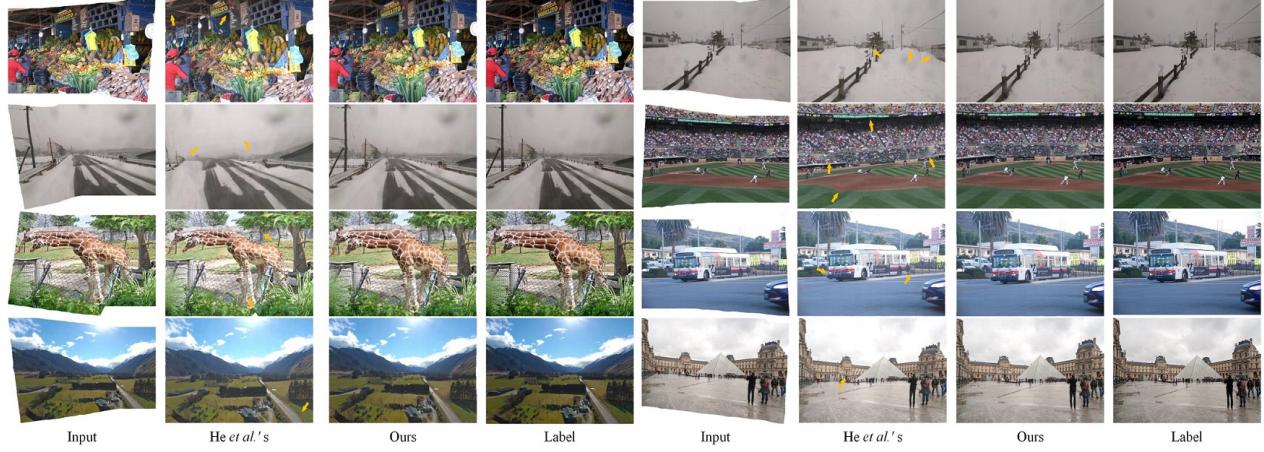
We conduct user studies on visual preference. Formally, we compare He *et al.*'s algorithm with our primary rectangling and final rectangling (as shown in Fig. 3) one by one. At each time, three images are shown on one screen: the input, He *et al.*'s rectangling, and ours (primary or final). We shuffle the order of different methods each time. The users may zoom in on the images and are required to answer which result is preferred. In this study, we invite 10 participants, including five researchers/students with computer vision backgrounds and five volunteers outside this community. The results are shown in Fig. 6, where our solution is preferred by more users.

### 5.2.4 Cross-Dataset Evaluation

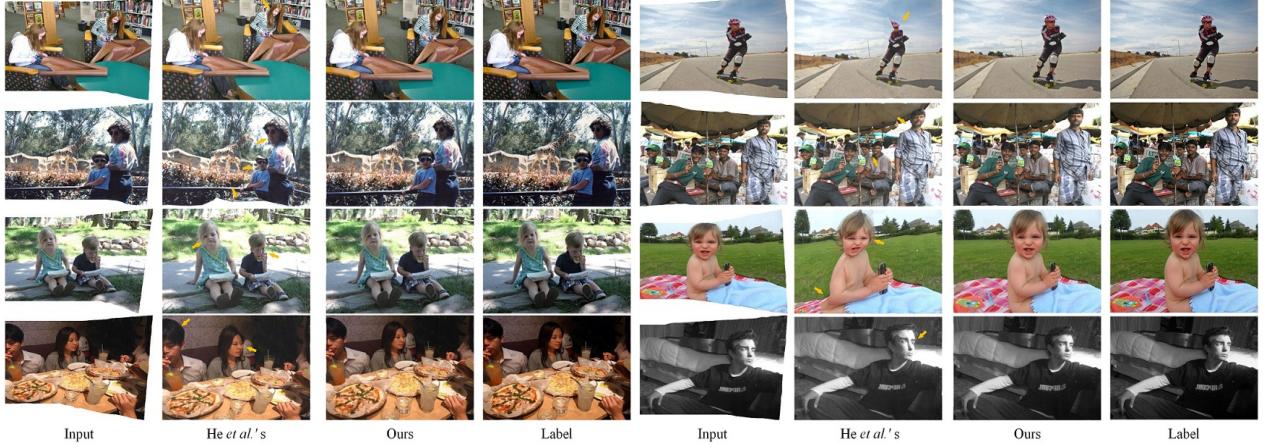
In this cross-dataset evaluation, we adopt the DIR-D dataset to train our model and test this model in other datasets.

Formally, we adopt different image stitching methods (SPW [18], LCP [10], and UDIS [24]) to stitch classic image stitching datasets [4,5,10,30]. Then, stitched images are used for rectangling using different algorithms. The results are shown in Fig. 8a, where our solution produces fewer distortions in rectangling results.

To show our effectiveness in more general scenes, where artifacts and projective distortions are not eliminated, we demonstrate rectangling results on a failure case of image stitching. As shown in Fig. 8b, our method still works well.



(a) Scenes with linear structures.



(b) Scenes with non-linear structures such as portraits.

Figure 7. Qualitative comparisons on DIR-D.

Table 3. Ablation studies on DIR-D.

	Loss function				Mesh resolution			Residual progressive regression			Metric		
	$\ell_c^a$	$\ell_c^p$	$\ell_b$	$\ell_m$	4x3	8x6	16x12	w/o	w/ (primary)	w/ (residual)	FID ↓	SSIM ↑	PSNR ↑
1	✓		✓			✓			✓		115.37	0.3498	14.43
2	✓	✓	✓	✓		✓			✓		24.57	0.6109	19.84
3	✓	✓	✓	✓	✓		✓		✓		22.43	0.6926	20.92
4	✓	✓	✓	✓	✓				✓		24.15	0.6361	20.16
5	✓	✓	✓	✓			✓		✓		22.32	0.6907	20.95
6	✓	✓	✓	✓			✓			✓	22.35	0.6902	20.93
7	✓	✓	✓	✓			✓			✓	<b>21.77</b>	<b>0.7141</b>	<b>21.28</b>

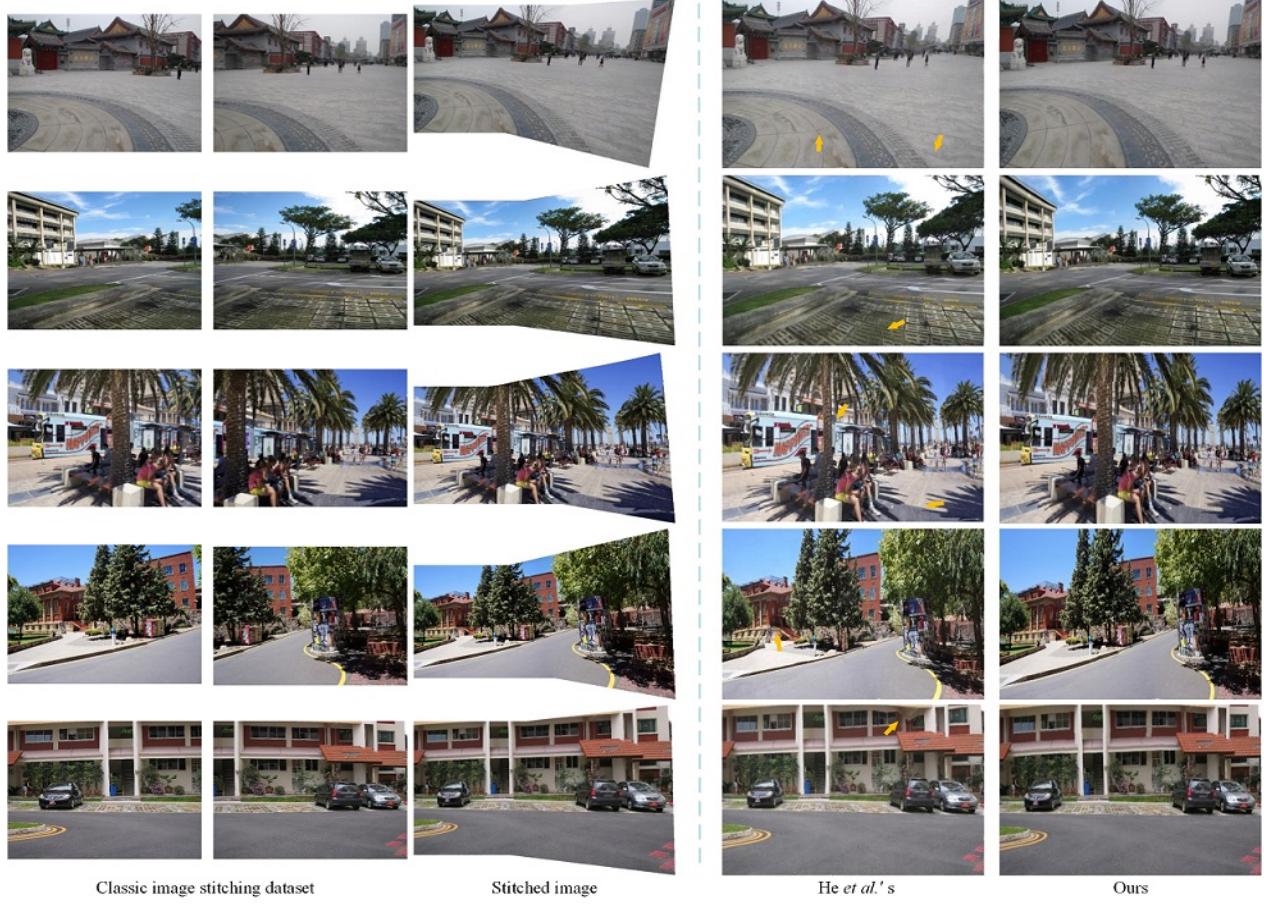
### 5.3. Ablation Studies

The proposed network is simple but effective. We validate the effectiveness of every module on DIR-D.

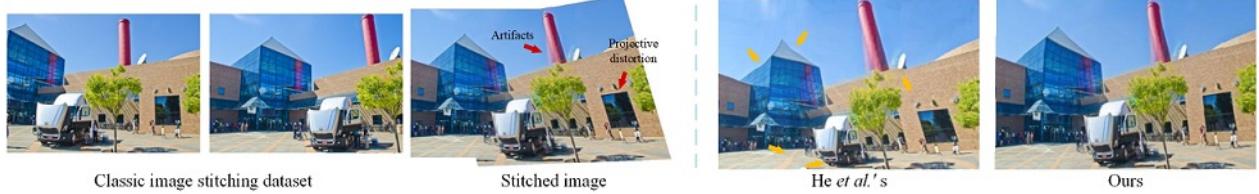
**Loss function.** We ablate the residual regressor as the baseline structure and evaluate the effectiveness of different constraint terms in our objective function. As shown in experiment 1-3 of Table 3, both the content term and mesh term can significantly improve our performance.

**Mesh resolution.** We test different mesh resolutions of  $4 \times 3$ ,  $8 \times 6$  and  $16 \times 12$ . As shown in the experiment 3-5 of Table 3,  $4 \times 3$  mesh decreases the rectangling performance while  $8 \times 6$  mesh and  $16 \times 12$  mesh give similar results. Nevertheless,  $16 \times 12$  mesh brings more computational costs, thus we adopt the  $8 \times 6$  mesh in our implementation.

**Residual progressive regression.** We validate the effectiveness of our residual progressive regression strategy



(a) Rectangling high-quality stitched images. The stitching datasets are from [4, 5, 30], and the stitching algorithms are from [10, 24].



(b) Rectangling low-quality stitched images, in which artifacts and distortions can be found. We adopt [24] to stitch images from [31]. We discuss this low-quality stitching example to show the effectiveness of our method in general real stitching scenes.

Figure 8. Cross-dataset qualitative comparisons. The arrows highlight the distorted regions.



Figure 9. Ablation study of the residual progressive regression strategy on other dataset ('potberry' [10]). The circles highlight the regions with uneven boundaries.

in experiment 6-7 of Table 3, where the residual regressor continues to refine the rectangling results based on the primary regressor. Although the improvement on the DIR-D

dataset is slight, this strategy enhances our generalization capability to avoid the uneven boundaries in other datasets as shown in Fig. 9.

## 6. Conclusion

In this paper, we propose the first deep image rectangling solution and dataset for image stitching. Compared with traditional two-stage methods, the proposed solution is a one-stage method, enabling efficient parallel computation with a predefined rigid target mesh. Besides, our solution can preserve both linear and non-linear structures, demonstrating the superiority over the existing methods both quantitatively and qualitatively.

## References

- [1] Shai Avidan and Ariel Shamir. Seam carving for content-aware image resizing. *ACM Trans. Graphics*, 26(3):10–es, 2007. 3
- [2] Che-Han Chang, Yoichi Sato, and Yung-Yu Chuang. Shape-preserving half-projective warps for image stitching. In *Proc. CVPR*, pages 3254–3261, 2014. 2
- [3] Yu-Sheng Chen and Yung-Yu Chuang. Natural image stitching with the global similarity prior. In *Proc. ECCV*, pages 186–201. Springer, 2016. 1
- [4] Junhong Gao, Seon Joo Kim, and Michael S Brown. Constructing image panoramas using dual-homography warping. In *Proc. CVPR*, pages 49–56, 2011. 6, 8
- [5] Junhong Gao, Yu Li, Tat-Jun Chin, and Michael S Brown. Seam-driven image stitching. In *Eurographics*, pages 45–48, 2013. 6, 8
- [6] Kaiming He, Huiwen Chang, and Jian Sun. Content-aware rotation. In *Proc. ICCV*, pages 553–560, 2013. 2
- [7] Kaiming He, Huiwen Chang, and Jian Sun. Rectangling panoramic images via warping. *ACM Trans. Graphics*, 32(4):1–10, 2013. 1, 2, 3, 4, 5, 6
- [8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proc. ICCV*, pages 2980–2988, 2017. 10
- [9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017. 6
- [10] Qi Jia, ZhengJun Li, Xin Fan, Haotian Zhao, Shiyu Teng, Xincheng Ye, and Longin Jan Latecki. Leveraging line-point consistence to preserve structures for wide parallax image stitching. In *Proc. CVPR*, pages 12186–12195, 2021. 2, 6, 8, 12
- [11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5
- [12] Wei-Sheng Lai, Orazio Gallo, Jinwei Gu, Deqing Sun, Ming-Hsuan Yang, and Jan Kautz. Video stitching for linear camera arrays. *arXiv preprint arXiv:1907.13622*, 2019. 2
- [13] Kyu-Yul Lee and Jae-Young Sim. Warping residual based image stitching for large parallax. In *Proc. CVPR*, pages 8198–8206, 2020. 1
- [14] Dongping Li, Kaiming He, Jian Sun, and Kun Zhou. A geodesic-preserving method for image warping. In *Proc. CVPR*, pages 213–221, 2015. 2, 3, 4
- [15] Jing Li, Baosong Deng, Rongfu Tang, Zhengming Wang, and Ye Yan. Local-adaptive image alignment based on triangular facet approximation. *IEEE Trans. on Image Processing*, 29:2356–2369, 2019. 2
- [16] Jing Li, Zhengming Wang, Shiming Lai, Yongping Zhai, and Maojun Zhang. Parallax-tolerant image stitching based on robust elastic warping. *IEEE Trans. on Multimedia*, 20(7):1672–1687, 2017. 5, 12
- [17] Nan Li, Yifang Xu, and Chao Wang. Quasi-homography warps in image stitching. *IEEE Trans. on Multimedia*, 20(6):1365–1375, 2017. 2
- [18] Tianli Liao and Nan Li. Single-perspective warps in natural image stitching. *IEEE Trans. on Image Processing*, 29:724–735, 2019. 2, 6, 12
- [19] Chung-Ching Lin, Sharathchandra U Pankanti, Karthikeyan Natesan Ramamurthy, and Aleksandr Y Aravkin. Adaptive as-natural-as-possible image stitching. In *Proc. CVPR*, pages 1155–1163, 2015. 2
- [20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. ECCV*, pages 740–755. Springer, 2014. 5
- [21] Wen-Yan Lin, Siying Liu, Yasuyuki Matsushita, Tian-Tsong Ng, and Loong-Fah Cheong. Smoothly varying affine stitching. In *Proc. CVPR*, pages 345–352. IEEE, 2011. 1
- [22] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a “completely blind” image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012. 6
- [23] Lang Nie, Chunyu Lin, Kang Liao, Shuaicheng Liu, and Yao Zhao. Depth-aware multi-grid deep homography estimation with contextual correlation. *IEEE Trans. on Circuits and Systems for Video Technology*, 2021. 3
- [24] Lang Nie, Chunyu Lin, Kang Liao, Shuaicheng Liu, and Yao Zhao. Unsupervised deep image stitching: Reconstructing stitched features to images. *IEEE Trans. on Image Processing*, 30:6184–6197, 2021. 1, 5, 6, 8
- [25] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4
- [26] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. *arXiv preprint arXiv:2109.07161*, 2021. 1, 2, 3, 12
- [27] Jing Tan, Shan Zhao, Pengfei Xiong, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical wide-angle portraits correction with deep structured models. In *Proc. CVPR*, pages 3497–3505, 2021. 2
- [28] Piotr Teterwak, Aaron Sarna, Dilip Krishnan, Aaron Maschinot, David Belanger, Ce Liu, and William T Freeman. Boundless: Generative adversarial networks for image extension. In *Proc. ICCV*, pages 10521–10530, 2019. 2, 3
- [29] N Venkatanath, D Praneeth, Maruthi Chandrasekhar Bh, Sumohana S Channappayya, and Swarup S Medasani. Blind image quality evaluation using perception based features. In *Twenty First National Conference on Communications*, pages 1–6, 2015. 6
- [30] Julio Zaragoza, Tat-Jun Chin, Michael S Brown, and David Suter. As-projective-as-possible image stitching with moving dlt. In *Proc. CVPR*, pages 2339–2346, 2013. 1, 6, 8
- [31] Fan Zhang and Feng Liu. Parallax-tolerant image stitching. In *Proc. CVPR*, pages 3262–3269, 2014. 8
- [32] Yun Zhang, Yu-Kun Lai, and Fang-Lue Zhang. Content-preserving image stitching with piecewise rectangular boundary constraints. *IEEE Trans. on Visualization and Computer Graphics*, 27(7):3198–3212, 2020. 3

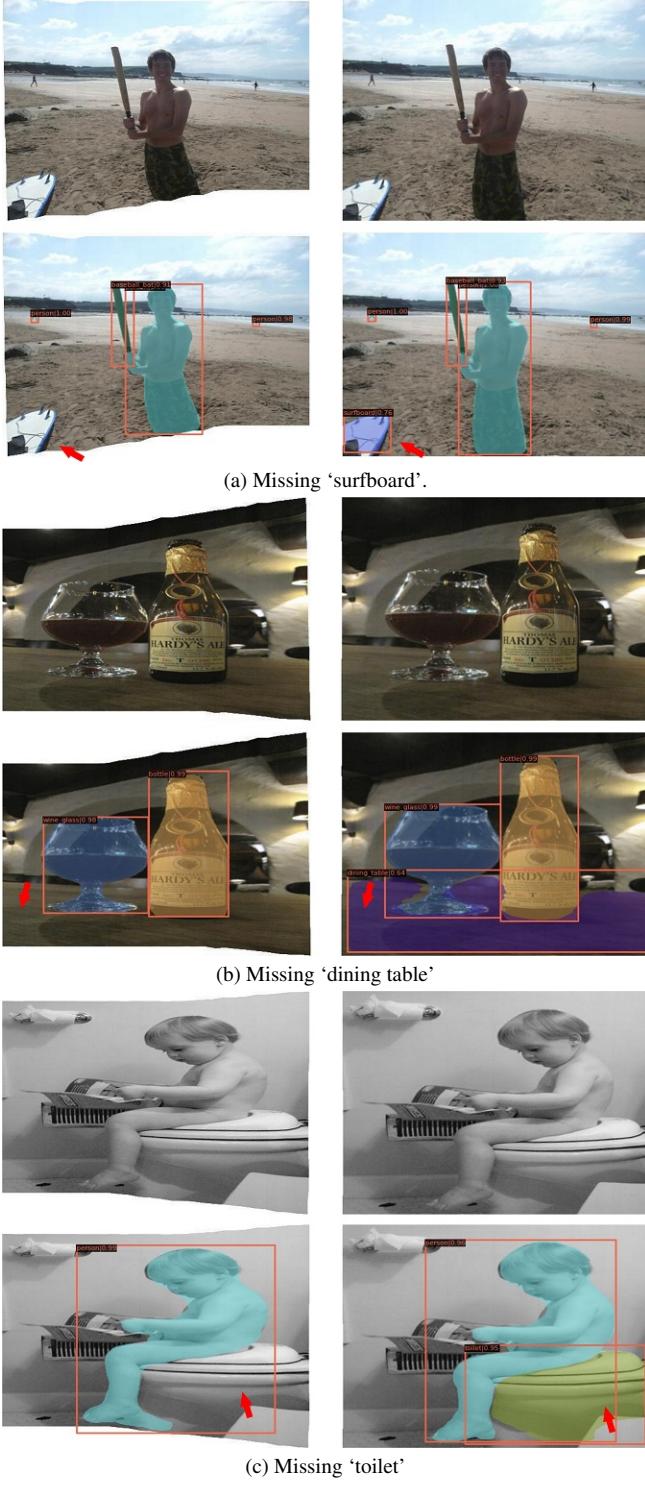


Figure 10. Object detection and semantic segmentation results of the stitched images and our rectangling results. The arrows highlight the missing parts.

## A. Overview

In the supplementary material, we first demonstrate the benefits of image rectangling for scene reasoning in Section **B**. Then, we illustrate more experimental results of our solution in Section **C**, including our rectangling results on the DIR-D dataset and other datasets.

## B. Benefits for Scene Reasoning

The proposed rectangling solution offers a nearly perfect visual perception for users by eliminating the irregular boundaries in image stitching. It can also help downstream vision tasks such as object detection and semantic segmentation, which is crucial for scene understanding. As shown in Fig. 10, the detection and segmentation results are derived from Mask R-CNN [8]. We can notice that the objects in the stitched images with irregular boundaries may be missing, such as the surfboard (Fig. 10a), the dining table (Fig. 10b), and the toilet (Fig. 10c). By contrast, our rectangling results ‘find’ the missing objects. We summarize the improvement as follows:

Almost all existing deep learning models (detection and segmentation) are trained on rectangular images, making them not robust to the regions around the boundaries in stitched images.

## C. More Results

More results on DIR-D are exhibited in Fig. 11, where our solution can deal with variable irregular boundaries and yield perceptually natural rectangular results.

Besides, more cross-dataset results are displayed in Fig. 12, which shows the superiority of rectangling over other solutions such as cropping and completion.



Figure 11. More results of our solution on DIR-D. Each triplet includes an input, a mask, and our rectangling result from left to right.



Figure 12. More cross-dataset results. The classic image stitching datasets ('crossline' [10], 'eyot' [10] and 'tower' [16]) are stitched by SPW [18]. We adopt LaMa [26] to complete the stitched images, and the rectangling results are generated by the proposed learning baseline.