Fast image reconstruction network in image stitching

XIE Menrui[1]* and SUN Bo[2]

1. Southeast University, Suzhou 215123, China

2. Quanzhou Institute of Equipment Manufacturing, Haixi Institutes, Chinese Academy of Sciences, Quanzhou 350108, China

(Received 7 March 2023; Revised 3 April 2023)

Compared with the traditional feature-based image stitching algorithm, the free-view image stitching algorithm based on deep learning has the advantages of fast stitching speed and good effect. However, these algorithms still cannot achieve real-time splicing speed. For the image reconstruction stage, we redesign a new fast image reconstruction network. This network is designed based on ShuffleNet, and the new network structure and loss function will reduce the time required for image reconstruction. In addition, this network can also reduce the performance loss after the network is lightweight. It is proved by experiments that the fast image reconstruction network can realize real-time high-resolution free-view image reconstruction.

Document code: A **Article ID:** 1673-1905(2023)10-0635-6

DOI https://doi.org/10.1007/s11801-023-3042-9

As a research content of computer vision, image stitching can be applied in many scenarios. In virtual reality technology[1], the application of image stitching can combine different scenes to build a larger virtual space. In remote sensing technology[2], image stitching can enable artificial satellites, aircraft and other equipment to collect more comprehensive, three-dimensional, and intuitive information scenes. In intelligent driving technology[3], the image information collected by the sensor can eliminate blind spots in the field of vision and reduce potential safety hazards through image stitching. In addition, image stitching technology can also be applied to medical treatment[4], monitoring[5], and so on. Therefore, realizing real-time image stitching on edge terminals has extremely high engineering application value. On edge terminals, such as vehicle-mounted computers and portable virtual reality devices, high-resolution real-time image stitching has important practical significance.

Traditional image stitching algorithms need to obtain external parameters through camera calibration to quickly calculate the deformation formula of the image, so as to achieve the real-time stitching effect of high-resolution images. In scenes without camera calibration conditions, the calculation of traditional image stitching algorithms is too large, and it is often necessary to reduce the calculation by reducing the resolution[2]. In this way, the cost of achieving real-time stitching effects usually results in a resolution that is too low to be practical. However, the increase in computing power brought about by the development of computer hardware has provided an important development foundation for deep learning. With the in-depth research and improvement of various deep learning algorithms and frameworks, deep learning has been applied to computer vision[6], date transmission[7], fiber spectrometer[8] and other aspects, and achieved ideal results in terms of accuracy and speed. In view of the advantages of deep learning over traditional image processing techniques, the combination of image stitching and deep learning can also bring performance and speed improvements. The existing image stitching based on deep learning[9], although it does not require camera calibration, can be flexibly applied to various scenes, and the stitching speed has also been greatly improved, it still cannot achieve real-time stitching at high resolution.

On the basis of the principle of the existing image stitching algorithm, we lighten the image reconstruction stage, redesign the network by combining the core ideas of ShuffleNet[10] and ResNet[11], reduce the parameters and calculation amount of the network model, and realize low computing resources under the real-time image reconstruction work. In addition, we also designed a new loss function to ensure that the image reconstruction effect after reducing the parameters and calculation amount will not have too much loss. Specifically, for a set of images that have been calibrated using homography or depth homography matrices, these images are first passed through several layers of lightweight convolutional layers to extract basic features. Then, through the processing of deep autoencoder, the overlapping regions are learned and the image is reconstructed. Finally, the idea of ResNet is used to connect the output layer of the

* E-mail: mendruxie@gmail.com

deep autoencoder to the input layer to reduce the loss of information in the convolution process to ensure the generation of high-resolution images. In addition, the new loss function can also reinforce the performance of the lightweight image reconstruction network through the comparison of image features.

The fixed-viewpoint image stitching algorithm is task-driven and designed for special application scenarios, such as autonomous driving and surveillance video. The conventional image stitching algorithm with a fixed viewing angle is to establish a parametric image alignment model through known camera calibration parameters when the relative positions of the two cameras are fixed, so as to achieve the image stitching effect. In fact, the homography transformation is the most common image alignment model, which uses methods such as translation, rotation, scaling, and vanishing point transformation to align one image to another. Moreover, the fixed camera position can ensure that after the complex initial alignment model parameters are calculated, the stitched image can be quickly obtained without a long calculation process. However, image stitching algorithms for fixed viewpoints are often end-to-end working models, which cannot be extended to stitched images of arbitrary views.

Due to the powerful feature extraction capabilities of convolutional neural networks (CNN), more and more image stitching projects choose to use CNN to achieve better stitching results. However, some of these methods only use CNN for the feature extraction stage[12], and some can only be used to stitch images of fixed viewpoints instead of free viewpoints[13]. In addition, there are more free-view image stitching algorithms based on deep learning that will be introduced in the next section.

In order to realize the image stitching of any view, the conventional free-view image stitching algorithm establishes the image alignment model through a feature-based method, and the conventional free-view image stitching algorithm is divided into two types based on different selections of target features. The first approach removes artifacts by aligning the target image with the reference image as much as possible. These methods divide the image into distinct regions and compute a homography for each distinct region. By applying a spatially varying warp on these regions, overlapping regions are well aligned and artifacts are significantly reduced. For example, double homography warping (DHW)[14] aligns the foreground and background of an image separately, which works well in scenes consisting of two main planes, but does not perform well in more complex scenes. The as-projective-as-possible (APAP)[15] method divides the image into dense grids, and each grid will assign a corresponding homography by weighting the features.

The second approach is by studying seam effects[16]. By optimizing the cost associated with the seam, the overlap can be divided into two complementary regions along the seam. Then, a stitched image is formed from the two regions. Feature-based solutions can significantly reduce artifacts in most scenes. Nevertheless, they still rely heavily on feature detection, so in scenes with few features or low resolution (in Fig.1), the stitching performance drops dramatically or even fails.

Fig.1 The Images that failed to stitch

Although the conventional free-view image stitching algorithm can reduce artifacts through various optimization methods to ensure stitching quality, more and more complex optimization algorithms will increase the time complexity of image stitching dramatically. Therefore, in order to balance the stitching image quality and stitching time, many researchers try to apply CNN to free-view image stitching. As mentioned earlier, the image stitching algorithm using CNN is either not a complete deep learning framework, or it can only stitch pictures with a fixed perspective. Then, a depth-of-view image stitching method[17] attempts to address these two issues. In this view-free solution, depth image stitching can be done by a depth homography module, a spatial transformer module and a depth image refinement module. However, all solutions are supervised methods, and since stitching labels are not available in real scenes currently, there are no real deep image stitching datasets. Therefore, these networks can only be trained on parallax-free synthetic datasets, resulting in unsatisfactory applications in real scenes.

To overcome the limitations of feature-based solutions and supervised deep solutions, an unsupervised deep image stitching framework[18] is proposed, which includes an unsupervised coarse image alignment stage and an unsupervised image reconstruction stage. At the same time, due to the use of real scenes as the dataset, the stitching quality and acceptable resolution of this framework are both enhanced compared to previous work. Nonetheless, like previous work on free-view image stitching, this work failed to reduce the stitching time of images to real-time effects.

As mentioned earlier, the currently best image stitching algorithm based on deep learning still cannot achieve real-time recognition speed. Therefore, after testing the algorithm, we found that lightweighting its image reconstruction stage can significantly reduce the stitching time. Currently, the image reconstruction stage used by this unsupervised deep image stitching framework is implemented through two branches: a low-resolution warping branch and a high-resolution refinement branch. In the case of building a network using CNN as the main layer, using low-resolution images as input in complex parts of the network can reduce the computational load and

inference time of the neural network. However, when stitching high-resolution images, information will be lost in the deformation branch, resulting in that the stitched image cannot restore the resolution of the input image. Therefore, we propose a new fast image reconstruction network, using the ShuffleNet block to replace the original convolutional layer, and directly use the original image for processing instead of splitting into two branches. Compared with other lightweight network structures, ShuffleNet can better balance inference speed and accuracy, and reduce image reconstruction time while ensuring reconstruction quality as much as possible. In addition, the ShuffleNet blocks can also help the model optimize the memory reading speed and further improve the reconstruction speed.

The fast image reconstruction network is shown in Fig.2. After the two aligned input images are superimposed into the network, the feature map is added after a layer of convolution. Afterwards, the input image augmented with feature maps passes through a lightweight deep autoencoder composed of ShuffleNet blocks. In this deep autoencoder, the encoding and decoding layers of the same size will be directly connected to ensure feature reuse and information integrity. It is worth noting that this lightweight deep autoencoder is composed of three ShuffleNet blocks (in Fig.3). The number of feature maps of the ShuffleNet keep block remains unchanged and is used to replace the original convolutional layer. The number of feature maps of the ShuffleNet down block is doubled, but the size of the feature maps will be quartered to replace the downsample layer. The number of feature maps of the ShuffleNet up block is half of the original, but the size of the feature maps will be expanded by four times to replace the upsample layer. Depth separable convolution is used in all ShuffleNet blocks to ensure that the effect of the network will not be much worse after reducing the amount of network parameters and calculations. After the lightweight deep autoencoder, the picture has completed the basic splicing. After that, it is necessary to use some ShuffleNet keep blocks to refine the effect of the picture. Before refining the picture, it is necessary to connect the processed feature map with the original image, so that more details of the original image can be learned.

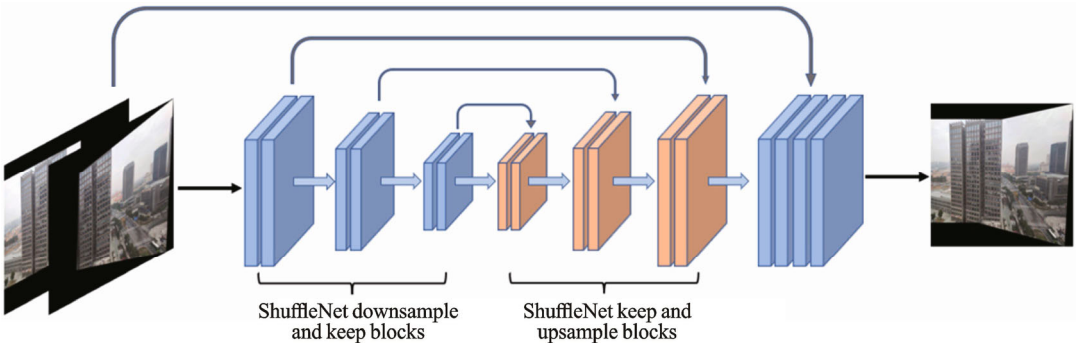

Fig.2 The structure of fast image construction network

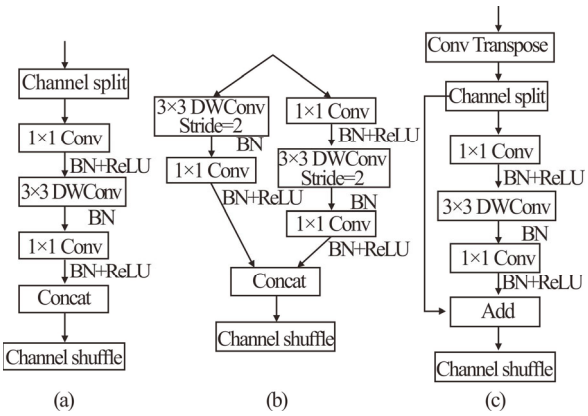

Fig.3 (a) ShuffleNet keep block; (b) ShuffleNet down block; (c) ShuffleNet up block

During the training of the model, the reconstruction rules for image stitching are learned through content masks and seam masks, which are generated in the same way as used in the original image reconstruction network mentioned earlier. Content mask can constrain the content of the stitched image to be consistent with the original image content. The seam mask can reduce the blur around the overlapping area, so that the overlapping area can transition naturally. The difference is that two content mask loss functions and one seam mask loss function are used in the fast image reconstruction network. Among them, the calculation formula of the loss function of the seam mask remains unchanged, and the calculation formulas of the loss function of the two content masks are shown as following equation

$$L_c^l = L_P\left(I_S \odot M_{AW}, I_{AW}\right), \tag{1}$$

$$L_c^h = L_P\left(I_S \odot M_{BW}, I_{BW}\right), \tag{2}$$

where L_c^l and L_c^h represent the low-level and high-level content mask loss functions. L_P is the perceptual loss[19]. I_S is the stitched picture. M_{AW} and M_{BW} are the content masks of the first warped image and the second warped image, respectively. I_{AW} and I_{BW} denote the first warped input image and the second warped input image, respectively. Specifically, in the perceptual loss, two kinds of content mask loss functions select the 'con3_3' layer and the 'conv5_3' layer in VGG19[20], respectively. By

comparing the feature differences between the stitched image and the original image in different layers by VGG19, the stitched image can be made with the original image as similar as possible in different perceptual characteristics. Thus, the total loss function of the entire network is shown in the following equation:

$$L = \lambda_S L_S + \lambda_c \left(L_c^l + L_c^h \right), \tag{3}$$

where λ_S and λ_c represent the weight parameters of the seam mask loss function and the content mask loss function, respectively, and L_S represents the loss function of the seam mask.

To train the model and verify its stitching performance and speed, we use a combined dataset, partly from variable motion videos[21] and partly from real datasets captured by ourselves. In particular, these data contain images of different scenes, different disparities, and different overlapping ratios. Diverse data can make the model more robust and practical. So far, we have 10 440 data for training the model and 1 126 data for testing. After training, the effect of stitching pictures using this model is shown in Fig.4. Among them, case (5) and case (6) belong to the real dataset, with large parallax and low overlap rate. Even so, the fast image reconstruction network still achieves good results.

Fig.4 The results of image stitching

In addition to intuitive picture results, data results (in Tab.1) obtained through statistical comparison are equally important. Overall, the algorithm using deep learning far outperforms the traditional algorithm in splicing success rate and speed. Specifically, the algorithm used in Ref.[18] still has the highest splicing success rate. But in terms of stitching speed, our algorithm can reduce the average time spent stitching pictures to 0.067 s. And in terms of success rate, compared with the highest 98.43%, our model can optimize the time to the real-time level with only limited concessions.

Tab.1 Performance of different image stitching algorithms (including traditional algorithms, supervised learning and unsupervised learning)

Method	Success rate	Speed (average)
APAP[15]	92.56%	38.62 s
Robust ELA[16]	87.92%	25.77 s
Ref.[18]	**98.43%**	0.47 s
EPISNet[22]	96.81%	2.83 s
Ours	97.78%	**0.067 s**

Comparing the intuitive stitching results of different methods (in Fig.5), it can be found that all methods achieve successful stitching results in small baseline images. In the dark night scene, APAP, robust ELA, and EPISNet all have obvious transition lines at the edge of the overlapping area. In the large baseline scene, EPISNet produces artifacts and robust ELA appears blurred. Overall, APAP, Ref.[18] and our method all generate good stitched images.

Looking specifically at the splicing results, it can be found that in Fig.6, the spliced images did not fail due to the reduction in resolution. Even in very challenging dark night scenes, the fast image reconstruction network can still achieve good results. Therefore, the resolution of the image to be stitched does not have a great influence on the reconstruction success rate.

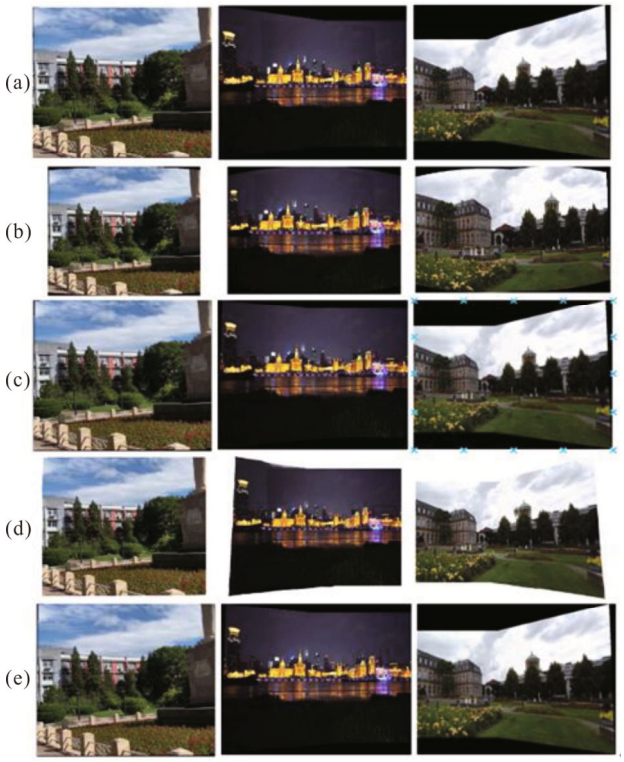

Fig.5 The results of different methods order by row: (a) APAP; (b) Robust ELA; (c) Ref.[18]; (d) EPISNet; (e) Ours

Fig.6 The results of different resolutions

This paper proposes a network model that can be quickly derived during the reconstruction phase of image stitching. This network model can complete the reconstruction work in stitching the aligned free-view images. By using a new network structure based on ShuffleNet and an optimized loss function, the time required to stitch pictures can be greatly reduced at the expense of a small amount of success rate. However, the experimental results show that the error correction ability of the fast reconstruction network needs to be improved for data that has errors in the image alignment stage. At the same time, the performance of the network on close-range images is not satisfactory. To solve these problems, better image alignment models can be considered. Apart from this, replace the parameters of the perception function in the loss function to reduce the dependence on the image alignment model. In addition, the time spent on splicing can be further reduced through the model lightweight method in the deployment stage.

Ethics declarations

Conflicts of interest

The authors declare no conflict of interest.

References

[1] ANDERSON R, GALLUP D, BARRON J T, et al. Jump：virtual reality video[J]. ACM transactions on graphics (TOG), 2016, 35(6)：1-13.

[2] LI J, ZHAO Y, YE W, et al. Attentive deep stitching and quality assessment for 360° omnidirectional images[J]. IEEE journal of selected topics in signal processing, 2019, 14(1)：209-221.

[3] WANG L, YU W, LI B. Multi-scenes image stitching based on autonomous driving[C]//2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), June 12-14, 2020, Chongqing, China. New York：IEEE, 2020, 1：694-698.

[4] LI D, HE Q, LIU C, et al. Medical image stitching using parallel sift detection and transformation fitting by particle swarm optimization[J]. Journal of medical imaging and health informatics, 2017, 7(6)：1139-1148.

[5] GADDAM V R, RIEGLER M, EG R, et al. Tiling in interactive panoramic video：approaches and evaluation[J]. IEEE transactions on multimedia, 2016, 18(9)：1819-1831.

[6] HE K, CHEN X, XIE S, et al. Masked autoencoders are scalable vision learners[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 19-20, 2022, New Orleans, USA. New York：IEEE, 2022：16000-16009.

[7] FENG F, GAN J A, NONG J, et al. Data transmission with up to 100 orbital angular momentum modes via commercial multi-mode fiber and parallel neural networks[J]. Optics express, 2022, 30(13)：23149-23162.

[8] FENG F, GAN J, CHEN P F, et al. AI-assisted spectrometer based on multi-mode optical fiber speckle patterns[J]. Optics communications, 2022, 522：128675.

[9] ZHANG F, LIU F. Casual stereoscopic panorama stitching[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, 2015, Boston, USA. New York：IEEE, 2015：2002-2010.

[10] ZHANG X, ZHOU X, LIN M, et al. Shufflenet：an extremely efficient convolutional neural network for mobile devices[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 18-22, 2018, Salt Lake City, USA. New York：IEEE, 2018：6848-6856.

[11] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 26-July 1, 2016, Las Vegas, USA. New York：IEEE, 2016：770-778.

[12] HOANG V D, TRAN D P, NHU N G, et al. Deep feature extraction for panoramic image stitching[C]//Intelligent Information and Database Systems：12th Asian Conference, ACIIDS 2020, March 23-26, 2020, Phuket, Thailand. Berlin, Heidelberg：Springer International Publishing, 2020：141-151.

[13] SHEN C, JI X, MIAO C. Real-time image stitching with convolutional neural networks[C]//2019 IEEE International Conference on Real-time Computing and Robotics (RCAR), August 6-11, 2019, Irkutsk, Russia. New York：IEEE, 2019：192-197.

[14] GAO J, KIM S J, BROWN M S. Constructing image panoramas using dual-homography warping[C]//IEEE Conference on Computer Vision and Pattern Recognition, June 20-25, 2011, Colorado Springs, CO, USA. New York：IEEE, 2011：49-56.

[15] ZARAGOZA J, CHIN T J, BROWN M S, et al. As-projective-as-possible image stitching with moving DLT[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2013, Portland, Oregon, USA. New York：IEEE, 2013：2339-2346.

[16] LI J, WANG Z, LAI S, et al. Parallax-tolerant image stitching based on robust elastic warping[J]. IEEE transactions on multimedia, 2017, 20(7)：1672-1687.

[17] NIE L, LIN C, LIAO K, et al. A view-free image stitching network based on global homography[J]. Journal of visual communication and image representation, 2020,

73：102950.

[18] NIE L, LIN C, LIAO K, et al. Unsupervised deep image stitching：reconstructing stitched features to images[J]. IEEE transactions on image processing, 2021, 30：6184-6197.

[19] JUSTIN J, ALEXANDRE A, LI F F. Perceptual losses for real-time style transfer and super-resolution[C]// 14th European Conference on Computer Vision (ECCV), October 11-14, 2016, Amsterdam, The Netherlands. Heidelberg：Springer International Publishing, 2016：694-711.

[20] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04) [2022-12-25]. https：//arxiv.org/abs/1409.1556.

[21] ZHANG J, WANG C, LIU S, et al. Content-aware unsupervised deep homography estimation[C]//16th European Conference on Computer Vision (ECCV), August 23-28, 2020, Glasgow, UK. Heidelberg：Springer International Publishing, 2020：653-669.

[22] NIE L, LIN C, LIAO K, et al. Learning edge-preserved image stitching from large-baseline deep homography[EB/OL]. (2020-12-11) [2022-12-25]. https：//arxiv.org/abs/2012.06194.