Journal of Computers Vol. 34 No. 3, June 2023, pp. 19-29
doi: 10.53106/199115992023063403002

Camera Tripod Removal Model in Panoramic Images Based on Generative Adversarial Networks

Jian Wu[1], Honghui Deng[1], Fei Cheng[2], Hongjun Wang[2*]

[1] Guang'an Vocational & Technical College, Guang'an 638000, Sichuan, China
gazywj@163.com

[2] School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 61756, Sichuan, China
wanghongjun@swjtu.edu.cn

Received 30 May 2022; Revised 2 September 2022; Accepted 28 October 2022

Abstract. There are often residual images of the camera tripod in panoramic images, which may reduce the image quality and deteriorate the post-processing speed. To address this problem, a camera tripod removal network (TRNet) based on generative adversarial network is proposed. As an end-to-end model, the generator is designed to include recognition and reconstruction branches, which reduce the number of parameters and improve the training efficiency by sharing the encoder and correspond to scaffold recognition and texture reconstruction respectively. The recognition branch based on the U-Net structure can effectively identify the tripod area, while the reconstruction branch can brilliantly reconstruct the texture details through an intermediate layer formed by stacking dilated convolution residual blocks. Furthermore, spectral normalized Markov discriminator and multiple combined loss function are adopted to promote global texture consistency and thus result in a better texture filling effect. Finally, a data set of 400 panoramic images is constructed and experimental results on this data set demonstrate the better repair ability of TRNet against other state-of-the-art methods.

Keywords: panoramic image, tripod removal, generative adversarial network, dilated convolution residual block

1 Introduction

At present, 3D technology can be mainly divided into two categories, one is 3D scene modeling and real-time rendering, and the other is panoramic technology which is a branch of virtual reality. 3D panorama is widely used in Web3D due to its characteristics of simplicity and practicality, which can utilize real images to establish a virtual environment. Specifically, panorama requires field photography. Then the captured panoramic photos are further processed, such as denoising, desensitization, etc. Finally, the processed images are stitched together to generate scenes.

Panorama technology strives for photo-level realism and scene-level 3D presentation, and at the same time, its low cost has attracted the attention of many researchers. Compared to 3D modeling techniques, panoramas do not have sophisticated interactive effects and virtual presentation power, but the techniques that emerge from them can also help with 3D modeling. For example, the 3D modeling technology can be used to restore the 3D shapes of various objects in real scenes, and then the restored panoramic images are utilized to map on the surface of objects, which allow the restored scenes to be more realistic.

Panoramic photo shooting generally requires an ultra-wide angle lens and tripod. Compared with ordinary cameras, the former has a wider field of view and higher overall imaging efficiency, and the latter plays a role in fixing the camera to avoid the problem of uneven image stitching caused by camera shake. However, at the same time, due to the existence of these two factors, images of the camera tripod will have different degrees of residue in the panoramic image, which affects the user experience significantly.

In the situation with a small number of images, the method of eliminating tripod in images is mainly manual by post-processing personnel through image processing software. This method has the best tripod removal quality, but is the most time-consuming and labor-intensive. While in some commercial applications, the way to deal with this type of problem is to use other images for hard masking, which is simple and straightforward, but the user experience is not ideal.

In response to the problems above, a framework based on generative adversarial network (GAN) [8] for tripod

* Corresponding Author

removal (TRNet) in panoramic images is proposed from the perspective of image restoration. For the generator, TRNet utilizes two branches of identification and reconstruction to remove camera tripod in an end-to-end manner. At the same time, a common encoder is used to share weights for efficiency. Finally, as for the discriminator and loss function, a spectral normalized Markov discriminator and a joint loss function are used to further optimize the removal effect. In this paper, the key research problems are to precisely find the position of camera tripod in the image and to naturally reconstruct pixel point of tripod position in the image. Major contributions of this paper are summarized as follows.

(1) A camera tripod removal network (TRNet) model for panoramic images is firstly proposed to remove the tripod in 3D images and to accelerate the post-processing speed. Even though, there are a few methods for camera tripod removing in the 2D images.

(2) The loss function of TRNet model is proposed in detail, and the structures of TRNet are designed step by step.

(3) Extensive experiments have been conducted on a real-word dataset, which demonstrates that TRNet has superior quality and performance than state-of-the-art methods.

The remainder of this paper is organized as follows. In Section 2, existing related works is reviewed. In Section 3, the detail of the proposed model, including loss function and structure are illustrated. Experimental results are described in Section 4. Finally, conclusions are drawn in Section 5.

2 Related Work

In this section, previous studies that related to our work will be presented, including objects removal, image restoration, and panoramic image processing methods.

The removal of objects in images has received a lot of attention and has a wide range of applications and research prospects in recent years [1-2]. For example, the digital compositor needs to erase the weave used during the filming of stunt shots to avoid malfunctioning. Some scholars have also applied artificial intelligence technology to film restoration, mainly for the old film picture spots, scratches, flicker, dithering, mold, tearing, noise flicker, dithering, mold, tearing, noise, etc.

Both image restoration and removal of objects in images can be categorized as natural image restoration in computer vision, while the tripod removal problem can be translated into a texture restoration problem in the tripod region. Up to now, many methods of image restoration have been proposed, which can be broadly classified into three categories, including structure-based [3], block-based [4], and network-based [6-7] methods.

Structure-based algorithms generally use geometric methods to repair missing parts in an image, which can well represent the structure in the image information. Wang et al. [3] used an outline generator instead of an edge generator, which was more suitable for the case where the corrupted image contains distinct objects. By introducing structural information, this approach produces a more intuitive and clearer repair results.

Block-based algorithms [4] typically take a random pixel point at the boundary of the area to be restored as the center, and then select a block based on the texture features of the image according to the center first. Then the best match block is searched for intact areas of the image that do not need to be repaired on a block-by-block basis. Finally the contents of this best match block are used to restore the area to be repaired. This method allows the block to repair the texture information of the image excellently, but does not maintain the structural consistency of the repaired area with the background well.

Pathak et al. [5] proposed to combine encoder-decoder and generative adversarial networks for image restoration first. The proposed algorithm used a Context Encoder-Decoder (CE) to generate the missing parts. Specifically, the encoder was used to extract the depth features of the image through a convolution layer, and the decoder was used to utilize the extracted depth features in a deconvolutional manner. The model is able to repair moderately large missing images well, but there is a pixel discontinuity on the boundary of the missing region.

Since GANs [9, 13-14] were proposed, well-performing network structures such as Deep Convolutional Generative Adversarial Networks (DCGAN) [10], WGAN-GP [11], and Shift-Net [12] have emerged, all of which are based on generative adversarial networks obtained by continuous optimization, allowing the models to obtain clearer structures and finer texture details. Furthermore, GAN is redesigned to reconstruct the pixel point of the image, which can precisely learn the texture around the tripod and naturally restore the image. This is the big difference between the proposed model and the state of the art methods. Then, template of tripod is designed to find the exact position in the image, which is also another difference between them.

3 Panoramic Image Datasets

Datasets of panoramic images are not common and need to be specifically collected and processed. Therefore, we traveled to 13 scenic spots in Sichuan province and took more than 6,000 panoramic photos, which were screened and processed to form a panoramic image dataset containing various environments and ground textures. Furthermore, 400 panoramic images named "Panaroma_400" were selected for the tripod removal task.

3.1 Panoramic Imaging Methods

Panoramic images are usually available in two forms, specifically are equirectangular and Cubemap. Equirectangular is a simple map projection that projects lines of longitude and latitude equidistantly onto a rectangular plane. The projection is very easy to construct, as it forms a grid of equal rectangles for subsequent processing.

Cubemap, another storage format for panoramic images, is a collection of six square images that represent reflections in the environment. However, the distribution of pixels projected at the corners of each side is still less uniform compared to the ideal case.

In this paper, to ensure the image quality, a super wide-angle camera is used to capture the panoramic image which is stored in an isometric column projection. Most of the panoramic images obtained have a resolution of 8640×4320 and a small number have a resolution of 7680×3840.

3.2 Tripod Diversity

Tripod can be divided into stable and portable types, and is selected according to the environment. The stable type bracket with gimbal handle will cover more pixel points; while the portable type is less stable and covers less pixel points.

3.3 Texture Diversity

The Panaroma_400 dataset contains a variety of floor textures such as masonry floor, concrete floor, tile floor, asphalt floor, and floor rubber floor, covering various environments commonly found in daily life.

A binary mask image (mask) was specially created for each image to indicate the tripod area. The hexahedral cut of the original panoramic image produces six mapped images named front (u), back (b), left (l), right (r), top (u), and bottom (d), and the mask image that indicates the position of the camera mount in the bottom (d) image. The mask image with a resolution of 8640×4320 has a resolution of 2750×2750, and the panoramic image with a resolution of 7680×3840 corresponds to a mask image with a resolution of 2445×2445.

4 Model Structure Description

In this section, TRNet model is designed in detail, which includes three important parts, namely loss function, generator and discriminator. Loss function is the objective of TRNet model, and then generator is to find the camera tripod and reconstruct the content of camera tripod region, at last discriminator is to judge the quality of reconstruction.

4.1 Loss Function

In order to guide the goal of TRNet model, the loss function is designed to control the direction of these operations in the model. The loss function of TRNet consists of four components, which is calculated as the weighted sum of the reconstruction loss (L_{Rec}), content loss($L_{Content}$), mask loss(L_{mask}), and adversarial loss(L_{adv}), specifically are

$$L = \theta_1 L_{Rec} + \theta_2 L_{Content} + \theta_3 L_{mask} + \theta_4 L_{adv}. \quad (1)$$

L_{Rec} is the reconstruction loss which guarantees effective features are learned and can be represented as the function of real image (I_{gt}), standard mask image (M_{gt}) and other regions (I) as follows

$$L_{Rec} = \lambda_R \left\| (I_{gt} - I_R) \times M_{gt} \right\|_1 + \beta_R \left\| (I_{gt} - I_R) \times (1 - M_{gt}) \right\|_1. \quad (2)$$

The residuals from the real image are computed for the bracket region and the other regions separately. Since we pay more attention to the texture of the scaffold region, the value of λ_R is larger than β_R. Furthermore, in order to better learn the ground texture, in addition to the reconstruction loss mentioned above, TRNet also incorporates a content loss consisting of a combination of perceptual loss and style loss [15, 17-19],

$$L_{Content} = \lambda_s \sum_i^N L_{S_i} + \lambda_p L_{Perc}. \quad (3)$$

Instead of simply computing the difference between the output of the reconstructed branch and the real image, the perceptual loss is computed as the difference between their multi-layer feature mappings, which can be expressed as

$$I_m = I \odot (1 - M) + I_R \odot M, \quad (4)$$

$$L_{Perc} = \sum_n^N \left\| \phi_n(I_R) - \phi_n(I_{gt}) \right\|_1 + \sum_n^N \left\| \phi_n(I_m) - \phi_n(I_{gt}) \right\|_1. \quad (5)$$

I_m is the original image with the tripod region removed, while $\phi_n()$ denotes the feature mapping of the nth pooling layer (1, 2, 3) of the pre-trained model VGG-16.

Similar to the perceptual loss, the style loss constructs the Gram matrix from the multilayer feature mapping, except that it focuses more on the visual perception of the scaffold region, which is represented in the following form

$$G_i = (\phi_n(I_i))^T \cdot (\phi_n(I_i)), \quad (6)$$

$$G_{gt} = (\phi_n(I_{gt}))^T (\phi_n(I_{gt})), \quad (7)$$

$$L_{S_i} = \sum_n^N \frac{1}{H_n W_n C_n} \left\| G_i - G_{gt} \right\|_1. \quad (8)$$

In addition to reconstruction loss and content loss, mask loss and adversarial loss are also incorporated into the loss function to enhance the sharpness of the texture whose calculation are described in detail in Section 4.2 Generator and Section 4.3 Discriminator, respectively.

4.2 Generator

There are two steps to eliminate the camera tripod. The first step is to find the area of the tripod, and the second step is to reconstruct the area of the tripod. The quality of reconstruction is the degree of coincidence between the tripod area and the surrounding area. Therefore, the generator of TRNet is designed as a two-branch full convolutional neural network (FCN) [13] based on U-Net [16] , as shown in Fig. 1.

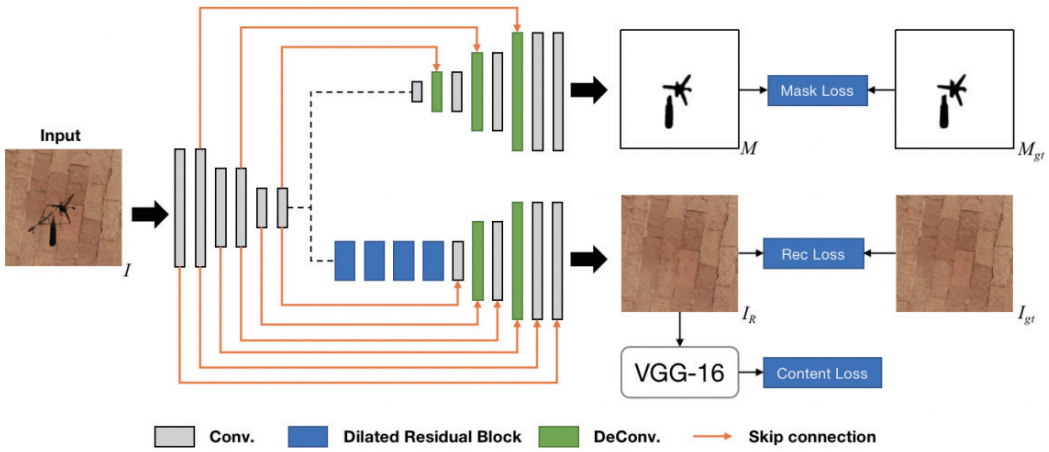


Fig. 1. Generator of TRNet

In the encoder stage, six convolutional layers are stacked to transform a tensor of size (3, 512, 512) representing the original image into a tensor of size (256, 64, 64). After the encoder, the network is divided into two decoder branches.

The recognition branch and the encoder form a typical U-Net image segmentation network, in which the output of the corresponding layer of the encoder stage is connected to the corresponding layer of the recognition branch by a jump connection, and a predicted mask image M is obtained using multi-layer de-convolution. Based on the above elaboration, the calculation formula of the mask loss can be obtained as

$$L_{mask} = 1 - Dice(M, M_{gt}), \tag{9}$$

$$Dice(M, M_{gt}) = \frac{2\left|M \cap M_{gt}\right|}{\left|M\right| + \left|M_{gt}\right|}, \tag{10}$$

$$\left|M\right| = \sum_i m_i \ , \tag{11}$$

$$\left|M_{gt}\right| = \sum_i m_i^{gt} \ , \tag{12}$$

$$\left|M \cap M_{gt}\right| = \sum_i m_i m_i^{gt} . \tag{13}$$

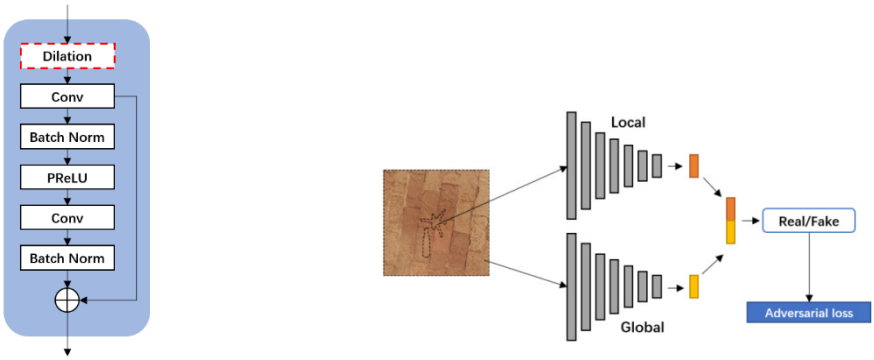
For the reconstruction branch, it shares the encoder with the recognition branch, reducing the number of weights, and at the same time, the capability of the recognition tripod learned by the recognition branch is passed to the reconstruction branch through the shared encoder. In this way, the trained TRNet does not need to input the mask image indicating the scaffold region when restoring photos, further improving the applicability.

In addition, residual blocks are introduced into the reconstruction branch to solve the problem of gradient disappearance due to excessive convolution. However, the use of residual blocks alone is not sufficient and its ability to extract information is still lacking. Dilated convolution can expand the perceptual field and extract a wider range of features without increasing the computational effort, which helps to reconstruct textures.

Therefore, in this paper, a new structure, the dilated residual block, is proposed, and its structure is shown in Fig. 2(a). Distinguished from the traditional residual block, a layer of dilated convolution is added in it. In the reconstruction branch, four blocks of cavity residuals with dilation values of 2, 4, 8, and 2 are stacked to form a special intermediate layer. This intermediate layer both avoids the degradation problem of deep neural networks

and takes advantage of the dilated convolution to improve the ability to extract features and repair the ground texture in a limited number of layers.

The final part of the reconstruction branch is the decoder. The decoder is designed as a symmetric structure to the shared encoder. At the same time, the output of the corresponding layer in the encoder is cascaded to the corresponding layer in the decoder by a jump connection, so that the reconstruction branch uses the fine-grained details learned in the encoder stage to construct the image in the decoder stage and obtain the output I_R of the reconstruction branch.



(a) Structure of the dilated residual block (b) Structure of the discriminator

Fig. 2. Structure of the dilated residual block and the discriminator

4.3 Discriminator

In this subsection, an worthy network of discriminator is designed to evaluate the performance of generator. In generative adversarial networks, the generator and discriminator components compete with each other with the aim of achieving Nash equilibrium through training. In TRNet, a spectrally normalized Markov discriminator [14] is employed to determine whether the output of the generator is true or false from both global and local perspectives to ensure the final output is of high quality, and its structure is shown in Fig. 2(b).

The discriminator stacks 7 convolutional layers with 4×4 kernel size and 2 steps to capture the features of Markovian patch. The final output of the discriminator is a patch feature of the shape H × W × C, where C is the number of channels. Then, these patches are penalized using hinge loss as adversarial loss to obtain the final probability that the input is a real or false scaffold erased image. Based on this, the adversarial loss can be calculated as follows

$$L_{adv} = L_D + L_G. \tag{14}$$

$$L_D = E_{x\sim P_{data}(x)}\left[ReLU(1-D(x))\right] + E_{z\sim Pz(z)}\left[ReLU(1+D(G(z)))\right]. \tag{15}$$

$$L_G = -E_{z\sim Pz(z)}\left[D(G(z))\right]. \tag{16}$$

5 Experimental Results and Analysis

In this section, the experimental data set, evaluation metrics, and experimental environment are briefly described. Subsequently, the results of ablation and comparison experiments are analyzed in detail to illustrate the performance of the proposed model.

5.1 Datasets

The dataset Panorama_400 contains 400 sets of images. To ensure the same data distribution, 448 sets of images were randomly selected as the training set and the remaining 50 sets of images were used as the test set. Since both the input and output images of TRNet have a resolution of 512×512, the high-resolution images in the Panorama_400 dataset are down-sampled to 512×512.

5.2 Evaluation Metric

Six evaluation metrics were used to fully validate the superiority of the model, namely Peak Signal to Noise Ratio (PSNR) [20-21], Structured Similarity Index Method (SSIM) [22], Mean Square Error (MSE), Average of the Grey level Absolute Difference (AGE) [23], the Percentage of Error Pixels between Two Images (PEPS) [23] and the Percentage of Clustered Error Pixels (PCEPS) [23].

The higher the value of PSNR and SSIM, the better the restoration effect. And smaller values of the four metrics, MSE, AGE, pEPs and pCEPS, indicate less difference between the two images, i.e., better restoration.

5.3 Ablation Study

Firstly, experiments related to the effectiveness of style loss were done, and the experimental results are shown in Table 1. The first and second rows of Table 1 indicate the evaluation results calculated with and without the use of style loss, respectively. The data show that the six metrics are significantly improved after using style loss, which proves that the style loss facilitates the texture transfer from the texture of the bracket region to the texture of the ground and helps the reconstruction of the texture of the bracket region.

In the reconstructed branch, an intermediate layer consisting of a stack of four dilated residual blocks is added, which exhibits a strong texture learning capability. In order to compare the variability between the proposed structure and other structures, five sets of comparison experiments were designed, the details are shown in Table 2. The above six results were tested and the results are shown in Table 3 and Fig. 3.

Table 1. The results of the ablation experiments on style loss, where SL denotes style loss

	PSNR	MSSIM (%)	MSE	AGE	PEPS	PCEPS
TRNet (IG) w SL	39.0425	97.7600	0.0002	1.0921	0.0045	0.0015
TRNet (IG) w/o SL	35.9096	92.0200	0.0004	1.4985	0.0064	0.0019

Table 2. Details of the 5 contrasting structures

Index	Details of structure
(a)	Intermediate layer consisting of 4 layers of dilated residual blocks
(b)	Intermediate layer consisting of 7 layers of residual blocks
(c)	Intermediate layer consisting of 21 layers of dilated convolution
(d)	Intermediate layer consisting of 21 layers of ordinary convolution
(e)	Intermediate layer consisting of 7 layers where all convolutions are replaced with residual blocks of the dilated convolution
(f)	Intermediate layer without any operation

Table 3. Results of multiple structures under different evaluation metrics

	PSNR	MSSIM (%)	MSE	AGE	PEPS	PCEPS
(a)	40.0321	97.5600	0.0002	0.9778	0.0045	0.0016
(b)	38.3456	95.2200	0.0003	1.0834	0.0057	0.0026
(c)	30.7756	73.9200	0.0013	5.1567	0.0499	0.0027
(d)	35.8676	94.3211	0.0006	1.5122	0.0072	0.0024
(e)	37.9347	95.3022	0.0003	1.1429	0.0056	0.0024
(f)	28.7644	64.8133	0.0020	6.4002	0.0734	0.0062

(f) is the shallowest network among the six structures, and there is a distinct stent image residue in its restored image, which proves that deeper networks are more conducive to texture reconstruction. Although (c) uses all cavity convolution with a larger receptive field, it ignores the influence of adjacent pixels, resulting in a more severe smearing in the restored area with obvious iridescence. (d) restored image is accurate in color, but when the ground texture is more complex, it will produce a more unconventional texture and the border transition is not natural enough. (b), compared with (d), avoids the problem of gradient disappearance through jump connection and activation function, so the overall performance is better and restores the brick joints and brick textures more accurately. Similarly, (e) also has more jump connections and activation functions compared with (c), and the restoration ability is also improved.

The experimental results show that the images restored using (a) produce less noise and accurately identify the location of the brick joints and the texture of the ground. Even on floors with more complex and variable textures, the cavity residual block accurately restores details with natural boundary transitions and disappearance of bracket contours. In terms of evaluation index, the cavity residual block also has certain advantages in comparison with a variety of structures, which can also be visually confirmed. There are two reasons the proposed model work so better than the existing methods. First, the template of camera tripod is designed to find the exact position in the image, which can increase the accuracy. Second, the discriminator of proposed model can learning the texture around camera tripod in the image, then it can help the algorithm reconstruct the texture on the pot of camera tripod.

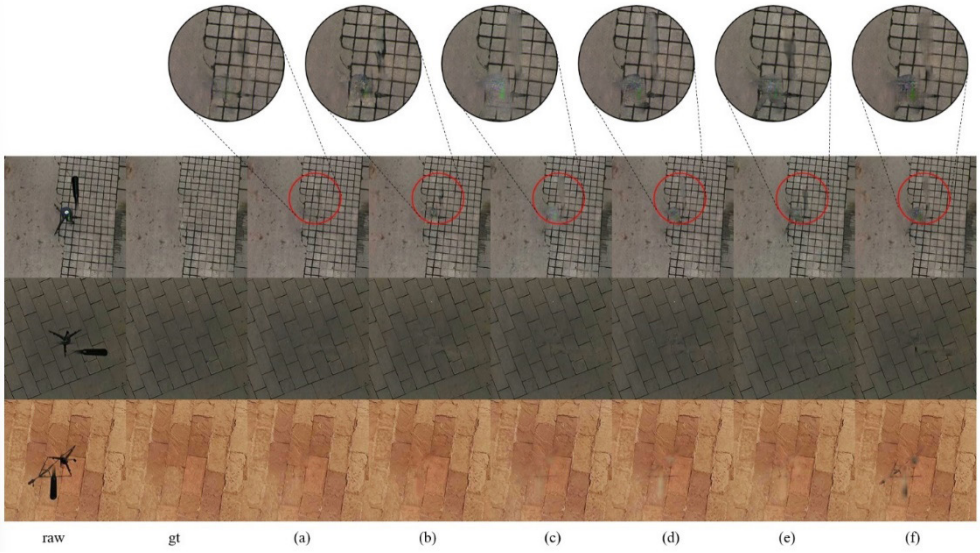


Fig. 3. Comparison of experimental results for multiple structures

(Raw is the original image and gt is the standard image.)

5.4 Comparison with Advanced Methods

In order to verify the superiority of the proposed model in this paper, experiments were conducted with other four advanced comparison algorithms under six evaluation metrics. The four compared algorithms are LBAM [24], EraseNet [13], CycleGAN [25] and EnsNet [23], and the experimental results are shown in Table 4 and Fig. 4.

As shown in Fig. 4, CycleGAN is able to roughly erase the camera tripod in the image, but because the ground has multiple complex textures, it may misjudge during the training phase, resulting in poor erasure results.

EraseNet has a good erasing effect in most scenes, but it does not reconstruct the texture within the tripod area well, resulting in a visual smearing effect.

Table 4. Data comparison with multiple advanced methods

	PSNR	MSSIM (%)	MSE	AGE	PEPS	PCEPS
CycleGAN	32.3421	91.2100	0.0011	2.7898	0.0182	0.0121
EraseNet	37.5255	96.4200	0.0002	1.6036	0.0052	0.0017
LBAM	29.3942	89.5500	0.0025	4.3021	0.0332	0.0112
EnsNet	33.7292	94.8600	0.0004	1.9829	0.0092	0.0033
TRNet (ours)	40.7644	97.9700	0.0001	0.9441	0.0040	0.0011

In the results obtained by LBAM, the contours and textures of the stabilized tripod are still relatively clear. And in some of the samples, the color reproduction also has a certain gap with the standard image, and there are rainbow patterns.

The overall performance of EnsNet is relatively good, but the convolution operation identifies the noise as a feature when it encounters a large range of noise, resulting in light spots in the results.

The TRNet proposed in this paper performs better in terms of color, texture, and edge transition. Thanks to the dilated residual block, TRNet can extract texture features in a larger range and promote smooth transition of texture on edges. Furthermore, loss functions such as style loss allow TRNet to better mimic the texture details of masonry. In particular, where the ground texture complexity is high, the texture within the bracket area is closer to the overall visual effect.

From the perspective of quantitative metrics, TRNet also has excellent performance compared to other networks, as evidenced by its first place performance in all six metrics. In the Friedman test [26], the X_F^2 statistics and p-values of TRNet are 23.832 and 8.6×10^{-5}, respectively. So the original hypothesis was rejected, i.e., TRNet significantly outperformed the other comparison algorithms.

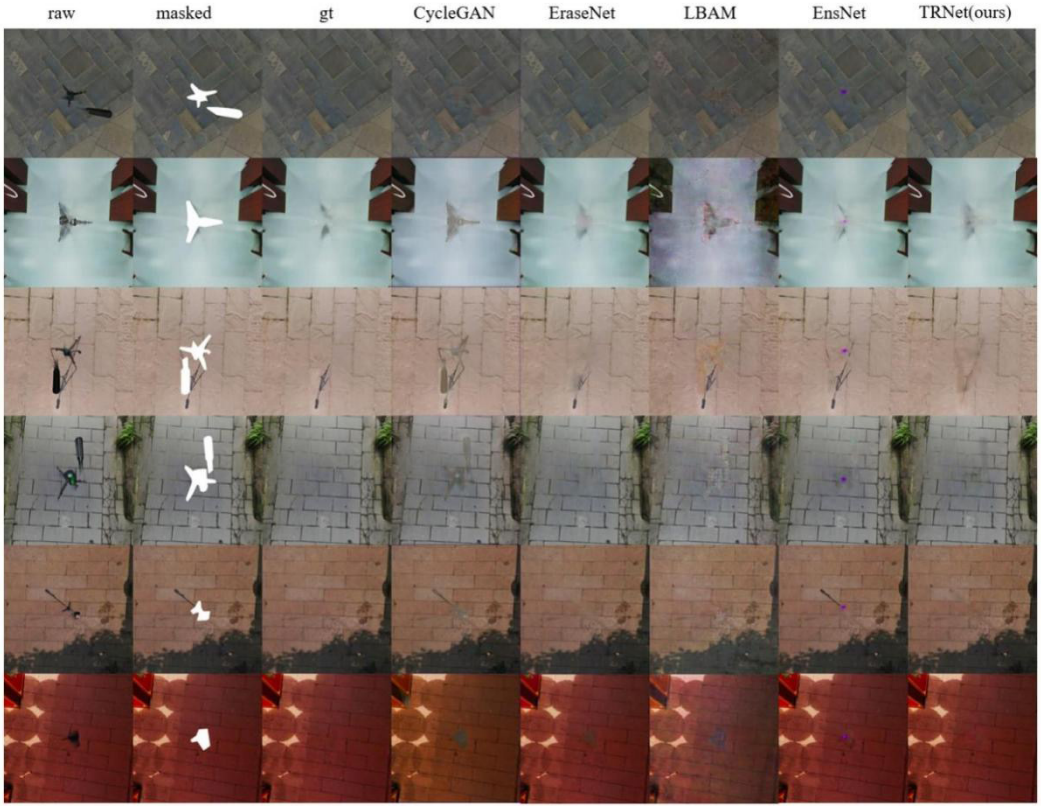


Fig. 4. The results of comparison with a variety of advanced methods

(Raw is the original image, masked is the mask image and gt is the standard image.)

6 Conclusion

In this paper, we propose a generative adversarial network-based camera bracket erasure model, which constructs generators consisting of recognition branches and reconstruction branches responsible for locating bracket regions and reconstructing tripod region textures from both image segmentation and image restoration, respectively. The recognition branch passes the learned tripod regions to the reconstruction branch through a shared encoder to achieve "end-to-end" tripod removal. To better reconstruct the texture, the residual blocks proposed in ResNet are optimized and the dilated residual blocks are proposed. Four cavity residual blocks are stacked in the reconstruction branch, which improves the ability to reconstruct ground texture and performs very well in the comparison of multiple structures. Experimental results on dataset Panaroma_400 show the brilliant recognition and texture reconstruction ability of the model proposed in this paper compared to other state-of-the-art models. The model is mainly an iterative algorithm, which requires more time in the the worst case. Therefore, future research on this model will mainly focus on improving time efficiency and parallel computing.

References

[1] C. Liu, Q. Gao, X. Wu, Exaggerated learning for clean-and-sharp image restoration, in: Proc. 2020 IEEE International Conference on Image Processing, 2020.
[2] L. Qi, J. Wu, X. Li, S. Zhang, S. Huang, Q. Feng, W. Chen, Photoacoustic Tomography image restoration with measured spatially variant point spread functions, IEEE Transactions on Medical Imaging 40(9)(2021) 2318-2328.
[3] W. Xiong, J. Yu, Z. Lin, J. Yang, X. Lu, C. Barnes, J. Luo, Foreground-aware image inpainting, in: Proc. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.
[4] V. Papyan, M. Elad, Multi-scale patch-based image restoration, IEEE Transactions on image processing 25(1)(2015) 249-261.
[5] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, A.A. Efros, Context Encoders: Feature Learning by Inpainting, in: Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[6] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, R. Timofte, Swinir: Image restoration using swin transformer, in: Proc. 2021 IEEE/CVF International Conference on Computer Vision, 2021.
[7] L. Chen, X. Lu, J. Zhang, X. Chu, C. Chen, HINet: Half instance normalization network for image restoration, in: Proc. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.
[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, Y. Bengio, Generative adversarial nets, in: Proc. 2014 Conference on Computer Vision and Pattern Recognition, 2014.
[9] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, T.S. Huang, Generative image inpainting with contextual attention, in: Proc. 2018 IEEE conference on computer vision and pattern recognition, 2018.
[10] P. Isola, J.Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proc. 2017 the IEEE conference on computer vision and pattern recognition, 2017.
[11] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A.C. Courville, Improved training of wasserstein gans, in: Proc. Advances in neural information processing systems 30 (NIPS 2017), 2017.
[12] Z. Yan, X. Li, M. Li, W. Zuo, S. Shan, Shift-net: Image inpainting via deep feature rearrangement, in: Proc. 2018 European conference on computer vision, 2018.
[13] C. Liu, Y. Liu, L. Jin, S. Zhang, C. Luo, Y. Wang, EraseNet: End-to-End Text Removal in the Wild, IEEE Transactions on Image Processing 29(2020) 8760-8775.
[14] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, T.S. Huang, Free-form image inpainting with gated convolution, in: Proc. 2019 IEEE/CVF International Conference on Computer Vision, 2019.
[15] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. 2016 IEEE conference on computer vision and pattern recognition, 2016.
[16] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Proc. 2015 International Conference on Medical image computing and computer-assisted intervention, 2015.
[17] J. Johnson, A. Alahi, F.-F. Li, Perceptual losses for real-time style transfer and super-resolution, in: Proc. 2016 European conference on computer vision, 2016.
[18] L.A. Gatys, A.S. Ecker, M. Bethge, A Neural Algorithm of Artistic Style, Journal of Vision 16(12)(2015) 326-326.
[19] L.A. Gatys, A.S. Ecker, M. Bethge, Image Style Transfer Using Convolutional Neural Networks, in: Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[20] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, H. Li, High-resolution image inpainting using multi-scale neural patch synthesis, in: Proc. 2017 IEEE conference on computer vision and pattern recognition, 2017.
[21] G. Liu, F.A. Reda, K.J. Shih, T.C. Wang, A. Tao, B. Catanzaro, Image inpainting for irregular holes using partial convolutions, in: Proc. 2018 European conference on computer vision, 2018.

[22] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13(4)(2004) 600-612.
[23] S. Zhang, Y. Liu, L. Jin, Y. Huang, S. Lai, Ensnet: Ensconce text in the wild, in: Proc. 2019 AAAI Conference on Artificial Intelligence, 2019.
[24] C. Xie, S. Liu, C. Li, M.M. Liu, W. Zuo, X. Liu, E. Ding, Image inpainting with learnable bidirectional attention maps, in: Proc. 2019 IEEE/CVF International Conference on Computer Vision, 2019.
[25] J.Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proc. 2017 IEEE international conference on computer vision, 2017.
[26] S. García, A. Fernández, J. Luengo, F. Herrera, Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power, Information sciences 180(10)(2010) 2044-2064.