IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 12, 2019, accepted December 1, 2019, date of publication December 6, 2019,
date of current version December 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2958111

360 Degree Panorama Synthesis From Sequential Views Based on Improved FC-Densenets

DANDAN ZHU[1], QIANGQIANG ZHOU[2], TIAN HAN[3], AND YONGQING CHEN[4]

[1]Artificial Intelligence Institute, Shanghai Jiao Tong University, Shanghai 200240, China
[2]School of Information and Computer, Shanghai Business School, Shanghai 201400, China
[3]Department of Computer Science, Stevens Institute of Technology, Hoboken, NJ 07030, USA
[4]Hainan Air Traffic Management Sub-Bureau, Haikou 570000, China

Corresponding author: Qiangqiang Zhou (zqqcsu@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61422112, Grant 61371146, Grant 61521062, and Grant 61527804.

ABSTRACT Inspired by the effectiveness of deep learning model, many panorama saliency prediction models based on deep learning began to emerge and achieved significant performance improvement. However, this kind of model requires a large number of labeled ground-truth data, and the existing panorama datasets are small-scale and difficult to train the deep learning models. To address this problem, we propose a novel panorama generative model for synthesizing realistic and sharp-looking panorama. In particular, our proposed panorama generative model consists of two sub-networks of generator and discriminator. At first, in order to make the synthesized panorama more realistic, we employ the improved Fully-Convolutional Densely Connected Convolutional Networks (FC-DenseNets) as the generator network. Secondly, we design a new correlation layer in the discriminator network, which can calculate the similarity between the generated image and the ground-truth image, and achieve the pixel level accuracy. The experimental results show that our proposed method outperforms other baseline work and has superior generalization ability to synthesize real-world data.

INDEX TERMS Virtual reality, panorama, saliency prediction, generative model, correlation layer.

I. INTRODUCTION

In recent years, with the rapid development of virtual reality (VR) and mobile Internet technology, the panoramic images have become increasingly popular. Compared to traditional 2D images, the 360^o panoramic images can capture the scene information in the range of $360^o \times 180^o$ in the horizontal and vertical viewing directions. By using head-mounted displays (HMD) such as HTC Vive, Samsung Gear VR and Oculus Rift, different perspective images can be rendered in real time, and viewers can view scenes in any direction to obtain an immersive VR experience. Based on these characteristics, 360^o panoramic images have received much attention and are widely used in many fields, such as entertainment [1], medical [2], education [3] and film and television [4], etc.

The resolution of 360^o panoramic images is several times that of conventional images, which makes it difficult to store and transmit panoramic images. However, the human visual attention mechanism has the ability to automatically select and allocate attention. When watching an image scene, the human eye can automatically process the region of interest (ROI) and selectively ignores other regions. Therefore, it is necessary to perform saliency prediction on the information in the panoramic images, in order to reasonably reduce the visual redundancy information. Saliency prediction on panoramic images not only improves the compression efficiency of panoramic images, but also reduces the transmission bandwidth.

In the past few years, deep learning technology has developed rapidly and has been applied in various realms, e.g., object detection, video summarization, image retrieval and person re-identification. The panoramic image saliency prediction model based on deep learning also began to emerge, and achieved better performance improvement. This kind of model requires a large number of manually labeled ground-truth data to train various proposed deep learning-based saliency prediction models, and the panoramic images are tested by the well-trained model to obtain the final predicted saliency map. However, the datasets

The associate editor coordinating the review of this manuscript and approving it for publication was Yongtao Hao.

TABLE 1. The properties of existing panoramic images and videos datasets.

Dataset	Scene	Images/videos	Number	Subjects	Ground-truth recorded
Sitzmann et al. [6]	Static	Image	22	169	Head and eye fixations
Rail et al. [7]	Static	Image	98	63	Head and eye fixations
Abreu et al. [8]	Static	Image	21	32	Head fixations
Upenik et al. [5]	Static	Image	104	40	Head fixations
Hu et al. [9]	Static	Image	70	27	Head fixations
Corbillon et al. [10]	Dynamic	Video	7	59	Head fixations
Lo et al. [11]	Dynamic	Video	10	50	Head fixations
Zhang et al. [12]	Dynamic	Video	104	27	Head and eye fixations
Xu et al. [13]	Dynamic	Video	208	31	Eye fixations
Ozcinar et al. [14]	Dynamic	Video	6	17	Head fixations
David et al. [15]	Dynamic	Video	19	57	Head and eye fixations
Deep 360 Pilot [16]	Dynamic	Video	342	5	Annotate salient object

for existing panoramic images and videos are small-scale and difficult to train the deep learning models. Table 1 summarizes the basic properties of the available panoramic images and videos datasets. As can be seen from Table 1, the largest panoramic image dataset is established by Upenik and Ebrahimi [5], which has the 104 panoramic images viewed by 40 subjects. However, this dataset only contained head fixations data that can be used without eye fixations data. Therefore, it is not feasible to utilize the existing panoramic image datasets to train the deep learning model.

In order to solve this problem, in this paper, we propose a novel panorama generative model via a generator and discriminator based on Fully-Convolutional Densely Connected Convolutional Networks (FC-DenseNets). Instead of Convolutional Neural Network (CNN), we apply the FC-DenseNets for panoramic images generation, which strengthens feature propagation, improves feature reuse, and obtains more accurate semantic features for generating panoramic images. In particular, when using conventional Generative Adversarial Network (GAN) to generate panoramic images, we notice that if the panoramic image contains highly complex background, the generated panoramic image is not clear or the original detail information is lost. To address these problems, we add the loss function based on local image patches to the proposed panorama generative model.

We briefly describe the implementation of the proposed panorama generative model. At first, we use a new perspective projection method to obtain overlapping left and right views, respectively. Secondly, the collected left and right views are used as input to the generator network and generate corresponding new left and right views, respectively. Then, we send the generated view and the counterpart ground-truth view into the discriminator network for image similarity discrimination. At last, all the generated views are seamlessly stitched to form a complete panoramic image.

To the best of our knowledge, the work we do to synthesize 360^o panoramic images from conventional images is currently rarely explored. The main contributions of this paper are summarized as follows.

- In order to help the synthesis process of panoramic images, we develop a new perspective projection method to obtain overlapping left and right views.
- We propose a novel panorama generative model, which consists of two sub-networks of generator and discriminator. In order to make the generated view more realistic, we adopt the improved FC-DenseNet structure. We design a new correlation layer in the discriminator network, which can calculate the similarity between the generated view and the ground-truth view to achieve the accuracy of per-pixel level.
- We conduct extensive experiments to show the effectiveness of the proposed model. Specially, our proposed generation model is capable of generating sharp-looking panoramic images without severe stitching artifacts or pixel inconsistencies.

II. RELATED WORK

In this section, we briefly review the approaches related to image generation.

A. TRADITIONAL IMAGE GENERATIVE MODEL

The generative model is an important model in probability statistics and machine learning. It represents a series of models for randomly generating observable data. The generative models are widely used and can be used to model different data, such as images, text, sound, etc. In this paper, we focus on the image generative model. Early research on image generative models mainly includes Gaussian Mixture Model (GMM) [17], Principle Component Analysis (PCA) [18] and Independent Component Analysis (ICA) [19]. These models can only model simple forms of data and cannot effectively model complex, irregular data distributions. In order to solve this problem, some well known models began to emerge, such as Markov Random Field (MRF) [20], Hidden Markov

Model (HMM) [21], Restricted Boltzmann Machine (RBM) [22], [23] and discriminant training generation models [24]. However, this kind of model requires limited application scenarios due to the lack of effective representation of features.

B. DEEP GENERATIVE MODEL

In recent years, with the rapid advancement of deep neural network (DNN) and the remarkable success in the computational domain, DNN has become the workhorse to solve various tasks (e.g. image classification, object detection, semantic segmentation, image retrieval). Because DNN can encode millions of parameters, it is more suitable for modeling complex data distributions, so that features can be effectively represented. There are many deep generative models [34], [35], all of which show promising results in generating images compared to the traditional generation models. Among these models, the more popular one is GAN model, which can generate a new image given the prior distribution of real data.

The GAN model shows excellent performance in generating low resolution images, but it does not perform well in generating high resolution images. The main reason for this problem is that the GAN model tend to BE instable during the training phase, and it is difficult to converge quickly. To address this problem, several generative models [25], [26] have been proposed recently. Wang *et al.* [25] propose a new high resolution image generative model that employs conditional Generative Adversarial Nets (cGAN) to conduct image synthesis at 2048 $\times$ 1024 resolution by inputting semantic labels. Despite the ability to generate high quality images, the drawback of this model is the need for semantic labels. Karras *et al.* [26] provide a model of training GAN in progressively growing manner and highly realistic generated images are obtained. Although this model can improve the quality of the sample, it is not generalized in practice, because it requires a lot of computing resources. Different from the previous methods, our method is inspired by the GAN model and we propose a novel model for synthesizing panoramic images. The model we propose not only generates high quality panoramic images, but also has excellent generalization in practice.

III. PROPOSED MODEL

In this section, we present the proposed panoramic image generative model in detail. Firstly, we apply the proposed perspective projection method to obtain multiple left and right views. Then, we propose a novel image generative model to synthesize new left and right views. After that, the generated views are seamlessly stitching to create a complete panoramic image. We will introduce them in details as follows.

A. LOCAL VIEW EXTRACTION

The conventional method of generating panoramic images is to place fisheye lens in four different directions to take four different images and then stitch them together.

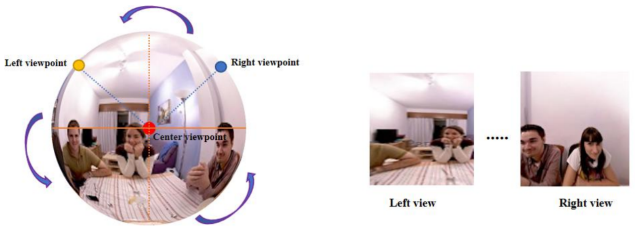


FIGURE 1. Diagram of local view extraction.

However, in contrast to the conventional approach, we project the input panoramic images to obtain multiple local left and right views. The input panoramic image has horizontal field of view (FoV) of 360 ^{o} and vertical FoV of 180 ^{o}. The specific projection method is shown in Fig. 1. At first, we map the input panoramic image to the sphere. Secondly, we place the camera in the center of the sphere (center viewpoint), and look at the left and right sides of the sphere through the center viewpoint and obtain the left and right viewpoint. Then we take corresponding viewpoint by rotating the camera to obtain multiple views. By this way, we collected 36 overlapping left and right views.

When we obtain the left and right views through this projection method, the next step is to generate new left and right views using the proposed generative model. Next we will introduce the network architecture of the proposed generative model.

B. NETWORK ARCHITECTURE

In this section, we focus on the architecture of the proposed panorama generative model. We propose the generative model, which consists of two deep neural network modules, namely the generator network and the discriminator network, respectively. Fig. 2 shows the architecture of our proposed model. In the generator work, we use the improved FC-DenseNet network structure as the backbone network. The input of this network is the original view, and the output is the corresponding synthetic new view. In the discriminator network, we design a new correlation layer for the similarity measure between the synthetic view and the ground-truth view.

1) GENERATOR NETWORK

The core part of our proposed generator network is composed of FC-DenseNet. FC-DenseNet is an extension of the DenseNet [27] that adds the upsampling path to make the output the same size as the input. The generator network we designed consist of 56 layers, including the downsampling path and the upsampling path. Specifically, the input to generator network is 224 $\times$ 224 view. The downsampling path of the FC-DenseNet is composed of 5 dense block (DB) layers and 5 transition down (TD) layers. The DB layer is composed of batch normalization [28], ReLU, 3 $\times$ 3 convolution layer and dropout layer. The transition layer is composed of batch normalization, followed by ReLU layer, a 1 $\times$ 1

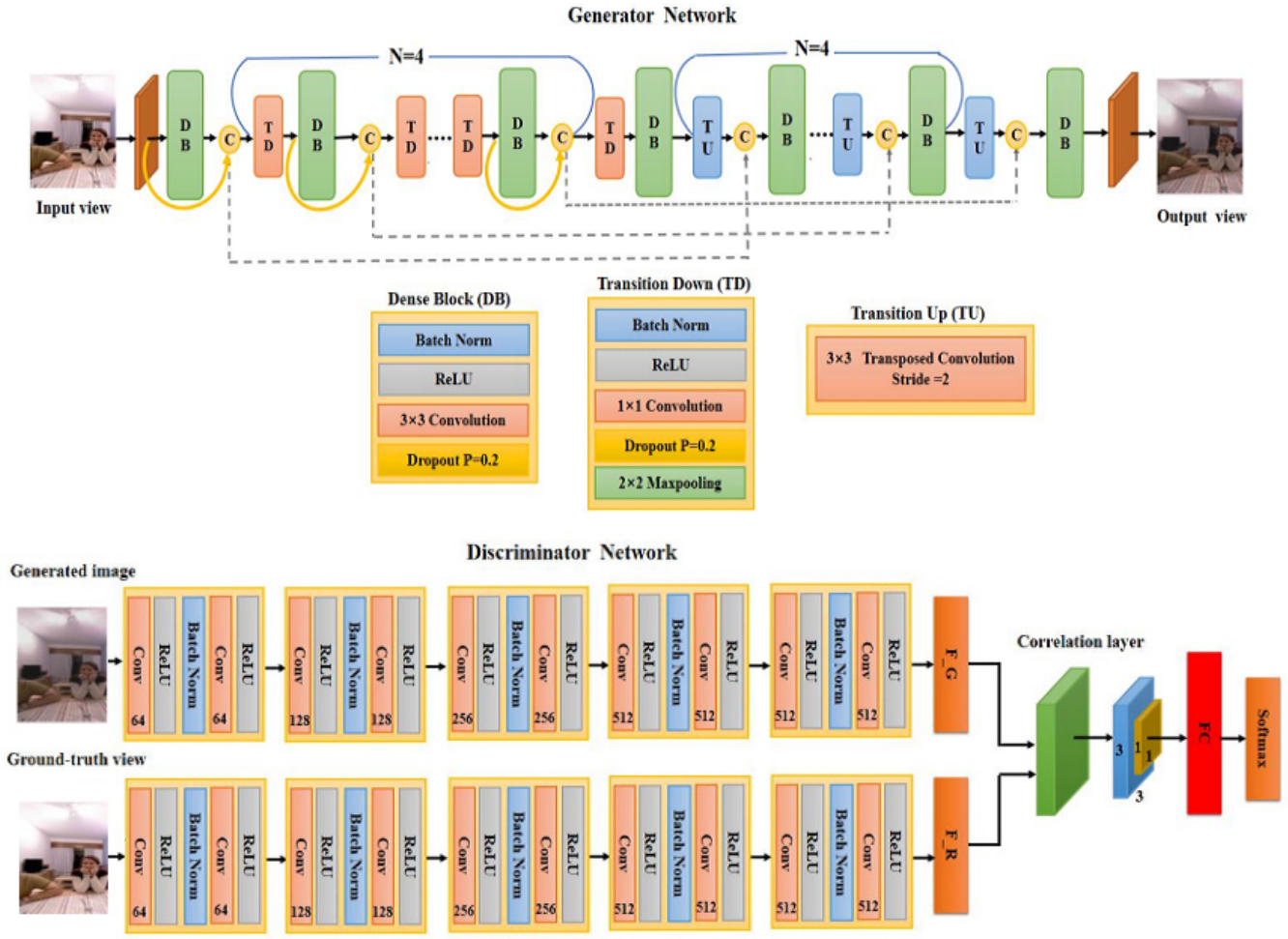


FIGURE 2. Architecture overview of our proposed panoramic images generation model.

convolutional layer (which keeps the number of feature maps untouched), dropout layer and a 2×2 maxpooling layer. The upsampling consist of 6 DB layers and 5 transition up (TU) layers. The TU layer is composed of a 3×3 transposed convolutional layer with stride 2 to solve the problem of reduced image resolution caused by the pooling operation in TD layer. It should be noted that the unsampled feature maps are concatenated to the ones with the same resolution from the downsampling path, so that the input of the new dense block can be formed. Different from the conventional networks, in order to prevent linear growth of the feature maps, we do not connect the input of the dense block to its output. On the last dense block, we use the transposed convolution to obtain the feature maps. The final layer of the generator network is a 1×1 convolutional layer, which outputs the generated new view.

2) DISCRIMINATOR NETWORK

In general, the discriminator network we propose is similar to the discriminator network structure in the conventional GAN model. However, different from the conventional GAN model, we design a novel layer (i.e. correlation layer) in the discriminator network, which can significantly improve the discriminative performance of the images. The architecture of the discriminator network is shown in Fig. 2. As can be seen from Fig. 2, our proposed discriminator network consists of two sub-networks, each of which consists of 5 tied convolutional blocks (convolutional layer, ReLU layer, batch normalization, convolutional layer and ReLU layer). The size of each convolution kernel in these two sub-networks is 3×3 and the stride is 1. The inputs to these two sub-networks are the synthetic view and the ground-truth view, respectively. In the process of discrimination, we measure the similarity between the synthetic view and the ground-truth view via a correlation layer. Specifically, we assume the synthesized view and the ground-truth view as V_s and V_g, respectively. The specific calculation process is to compare each patch in the V_s with each patch in the V_g, and the size of each patch is 3×3. For each position s_1 in the V_g, we calculate the correlation $C(s_1, s_2)$ in the neighborhood of the V_s centered at s_2 position. The size of the neighborhood is set to $M = 2d+1$, and $d = 1$ denotes the maximum displacement. The equation

is as follows:

$$C(s_1, s_2) = \sum_{d \in M} \langle V_g(s_1 + d), V_s(s_2 + d) \rangle. \quad (1)$$

As can be seen from Equation 1, we convolves data with other data. Unlike conventional convolution operations, data is convolved with spatial filter. By this calculation method, we can obtain the similarity measure between the synthetic view and the ground-truth view. If the C value is larger, it means that the synthetic view obtained by our generator network is more realistic. After the correlation layer, we add a 3×3 convolutional layer, a 1×1 max-pooling layer, and followed by a fully connected (FC) layer. The last layer of the discriminator network is the softmax function, which is used to output the probability that the generated view is a real view.

3) PANORAMA SYNTHESIS

Using GAN to generate realistic and clear panoramic images is a challenging task due to the instability of the training process. To address this problem, we propose a new view synthesis model. We can obtain a realistic view via the proposed model, but how to stitch the generated view into a complete panoramic image is still a difficult question. Specifically, we first need to find the same columns in different views (left and right view). Secondly, we resample the same columns in multiple views and seamlessly stitch them into a complete panoramic image.

C. OBJECTIVE FUNCTION

During the training process, we train a generator network and discriminator network simultaneously. The former inputs a noise vector z and outputs a generated image. The latter inputs a ground-truth image and generated image, and outputs a probability of distinguishing whether the sample is a ground-truth image or generated image. This training process is achieved by optimizing the adversarial loss function.

1) ADVERSARIAL LOSS

Ideally, the discriminator network needs to accurately determine whether the input image is a ground-truth image or generated image, and the generator network needs to do its best to deceive the discriminator network and let the discriminator network discriminate the generated image into ground-truth image. According to the description of the training process, we can define a loss function:

$$L_{adv}(G, D) = E[logD(I_m, I_n)] + E[log(1 - D(I_m, \hat{I}_n], \quad (2)$$

where G denotes the generator network, D represents the discriminator network, I_m is the input view, I_n is the ground-truth view and $\hat{I}_n$ is the view generated by the generator network G.

2) IMAGE PATCH LOSS

In the process of generating panoramic image using GAN, we find that if the image contains complex background and texture information, the generated panoramic image will be unclear and even some regions will be blurred. To solve this problem, we propose a weighting method based on image patches. The specific description is as follows: we first divide the image into three parts, namely foreground, middle and background patch; apply different weights to each part of the image. The specific loss function is expressed as follows:

$$L_{L1}(G) = E[\sum_{i=1}^{3} w_i || c_i(I_n) - c_i(G(I_m)], \quad (3)$$

where c_i refers to the i^{th} image patch, and w_i represents its corresponding weight.

3) FULL OBJECTIVE FUNCTION

The overall loss function of our proposed panoramic image generation model is defined as:

$$L = \arg \min_G \max_D [L_{adv}(G, D) + L_{L1}(G)], \quad (4)$$

which is the overall loss of our proposed GAN model training.

IV. EXPERIMENTS

In this section, we demonstrate the effectiveness of the proposed panorama generative model via experimental results. Firstly, we compare our method with other state-of-the-art methods on the benchmark datasets qualitatively and quantitatively. Secondly, in order to further show that our proposed model has superior generalization performance, we collect real-world data captured with smartphone and compare the results of synthetic panoramic images. At last, to show the superiority of using correlation layer to estimate the similarity between the synthetic view and the ground-truth view, we compare the proposed discriminator network with and without the correlation layer.

A. DATASET

To evaluate our proposed model, we use Salient360! [29] dataset. This dataset contains 60 images (360^o scenes), and eye tracking data provided as scan-paths and saliency maps and collected from 48 different observers. The dataset contains four different classes: indoor/outdoor natural scenes, scenes containing human faces, sports scenes and computer graphics contents. In order to ensure stability of the training, all images in the dataset are normalized to $[-1, 1]$.

B. EVALUATION METRIC

How to evaluate the performance of the generative model has not yet established a unified evaluation metric. Each evaluation metric has the advantage and disadvantages, and it is not able to evaluate the performance of generative model comprehensively and objectively. At present, the most commonly used evaluation metric for generative works [30] is Structural-Similarity (SSIM) and Peak Signal-to-Noise Ratio (PSNR). Therefore, we also employ SSIM and PSNR to evaluate our method in this paper.

SSIM This evaluation metric is used to measure the similarity in luminance, contrast and structure between

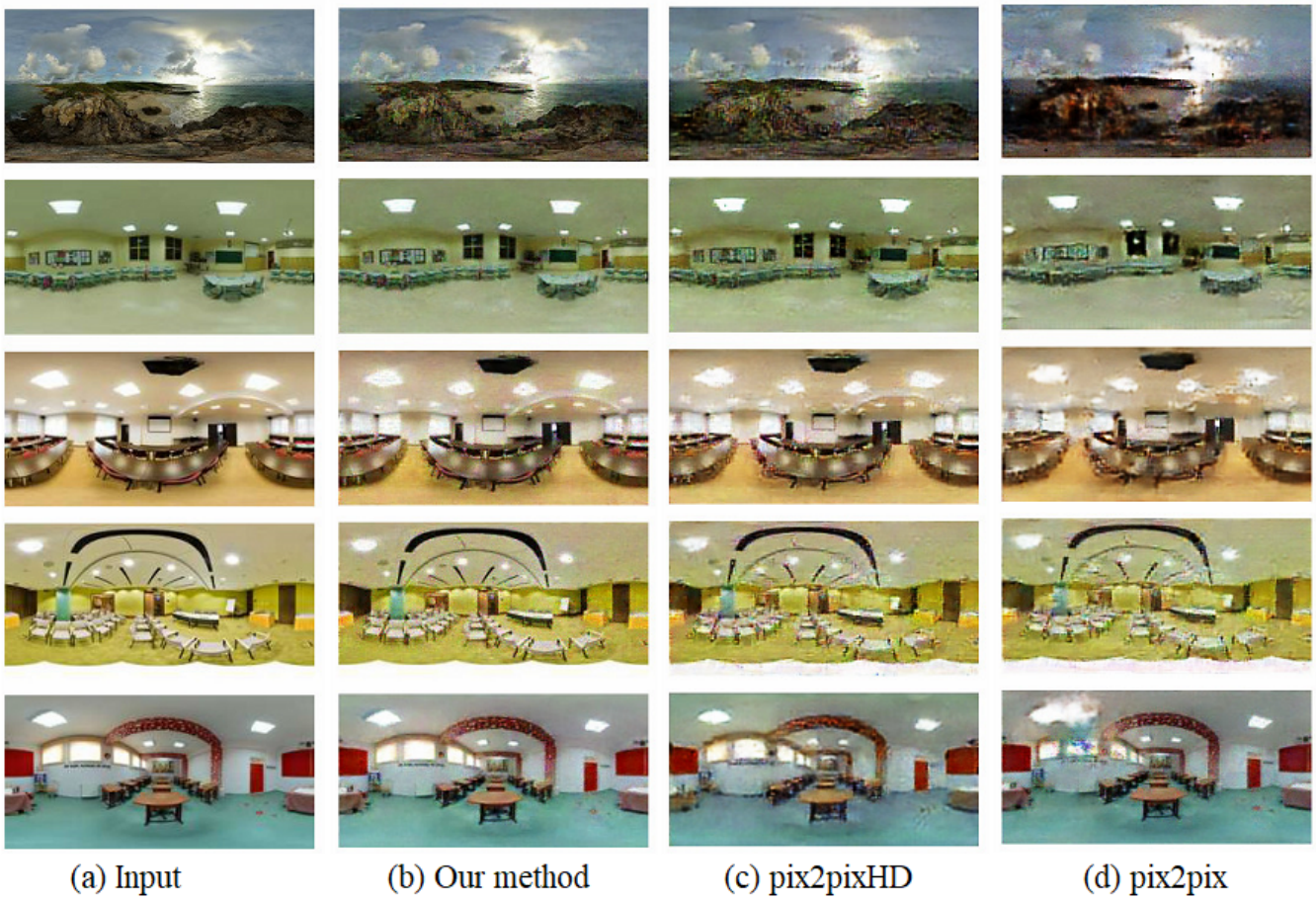


FIGURE 3. Qualitative comparison of our synthesized results with the baseline works on the Salient360! dataset.

two images. The value range of SSIM is between -1 and 1. If the value of SSIM is larger, it means that the similarity between the images being compared is higher. The corresponding equation is as follows:

$$SSIM(I_i, I_{i'}) = \frac{(2\mu_{I_i}\mu_{I_{i'}} + c_1)(2\sigma_{I_iI_{i'}} + c_2)}{(\mu_{I_i}^2 + \mu_{I_{i'}}^2 + c_1)(\sigma_{I_i}^2\sigma_{I_{i'}}^2 + c_2)}, \quad (5)$$

where I_i and $I_{i'}$ represent two images to be compared, μ_{I_i} and $\mu_{I_{i'}}$ denote the mean values of the image I_i and $I_{i'}$, respectively. σ_{I_i} and $\sigma_{I_{i'}}$ represent the standard deviations of the I_i and $I_{i'}$, and c_1 and c_2 are constants.

PSNR This is an objective metric widely used in evaluating image quality. In our paper, we compare the quality of the generated image with the quality of the ground-truth image. The higher the value of PSNR, the better the quality of the generated image and more realistic. Its calculation equation is as follows:

$$PSNR(I_i, I_{i'}) = 10log_{10}(\frac{MAX_{I_{i'}}^2}{MSE}), \quad (6)$$

where $MSE(I_i, I_{i'}) = \frac{1}{n}\sum_{j=1}^{n} ||I_i(j) - I_{i'}(j)||^2$. $MAX_{I_{i'}}$ represents the maximum value of the image color and $MAX_{I_{i'}} = 255$.

C. IMPLEMENTATION DETAILS

In the experiment, we implement our proposed model under the tensorflow [31] framework and train it by using Adam [32] with $\beta = 0.5$. The learning rate to train the proposed model is initialized as 0.00002 with batch size of 8. The entire model training takes about 10 hours. All the experiments are run on a workstation with NVIDIA GeForce RTX 2080Ti and 1TB RAM.

D. QUALITATIVE COMPARISON

Fig. 3 shows the visual comparison of our method and other two state-of-the-art methods for synthesizing panoramic images. Compared with the pix2pixHD [25] and pix2pix [33] methods, the panoramic images synthesized by our method are relatively smooth and sharp. Our method is different from the other two approaches mainly in the following two aspects: (1) the pix2pixHD method employs the semantic label as condition to synthesize the panoramic image, but it is difficult to obtain the semantic label in practice. (2) The pix2pix method requires a large number of pairs of images to conduct the image synthesis task during training. Although these baseline works perform well on semantic to image synthesis, image to image translation tasks, our method

FIGURE 4. Performance comparison of panorama generation for different methods using real-world data.

can successfully synthesize panoramic images without any requirements. As can be seen from Fig. 3, the images synthesized by these baseline works are relatively smooth, but they appear to be somewhat blurred, whereas the results obtained by our method are relatively sharp.

E. QUANTITATIVE COMPARISON

To further demonstrate the effectiveness of our proposed model, we present a quantitative comparison of the proposed model with other state-of-the-art method using SSIM and PSNR. Except for the generative methods, all other settings (e.g. dataset, hardware conditions, etc.) are kept untouched. The experimental results are shown in Table 2. As shown in Table 2, compared with several baseline works, our method obtains the best SSIM and PSNR values. Therefore, all these further demonstrate that our proposed panoramic image generative model achieves the best performance and can produce more realistic images.

TABLE 2. Comparison of SSIM and PSNR value for various methods.

Metric	Ours	pix2pixHD	pix2pix
SSIM	**0.4635**	0.3942	0.3522
PSNR	**16.3651**	15.1826	14.2190

F. SYNTHESIZE REAL-WORLD DATA

To further demonstrate the superior generalization ability of our proposed model, we set the smartphone to panoramic mode and then take some panoramic images for experimental verification. Specifically, we use 25 panoramic images taken with smartphone as testing set and employ our trained generative model to generate panoramic images. At the same time, for the fairness of comparison, we send the images taken from smartphone to pix2pixHD and pix2pix method for testing. The experimental results of the testing are shown in Fig. 4. From Fig. 4, we can clearly see that our proposed model can produce sharper panoramic images, while other baseline work to obtain relatively smooth and blurred panoramic images.

G. EVALUATION OF THE CORRELATION LAYER IN THE PANORAMA GENERATIVE MODEL

To validate the effectiveness of the proposed panorama generative model, we perform the performance comparisons for the proposed model with correlation layer and without correlation layer on the Salient360! dataset. Specially, in order to ensure the fairness of the comparison, all other setting of the discriminator network is kept untouched. We only add the correlation layer and remove the correlation layer to the experimental verification. We use SSIM to evaluate the role of the correlation layer in the discriminator network. The comparison results are shown in Table 3. It can be seen from Table 3, the SSIM values resulted from the panorama generative model with correlation layer is obviously better than the panorama generative model without correlation layer. This indicates that the correlation layer plays an effective refinement role onto improving patch matching performance.

TABLE 3. Performance comparisons of our proposed panorama generative model with correlation layer (with CL) and without correlation layer (without CL) on Salient360! dataset.

Metric	Our model with CL	Our model without CL
SSIM	**0.4635**	0.4257

V. CONCLUSION

In this paper, we present the work of synthesizing $360^o \times 180^o$ panoramic images from overlapping views. In order to help the panoramic image synthesis process, we first develop a new perspective projection method to obtain views (left and right). Secondly, we propose a novel panoramic image generative model, which consists of generator and discriminator network. In the generator network, we employ the improved FC-DenseNet architecture to synthesize more realistic views. We design a new correlation layer in the discriminator network to calculate the similarity between the generated view

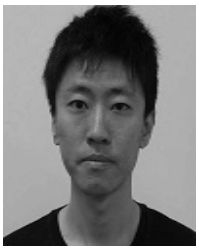
and the ground-truth view. The experimental results demonstrate the effectiveness of our proposed model and show that our model outperforms baseline work through qualitative and quantitative comparisons. To show the superior generalization performance of our proposed model, we use smartphone to collect the real scene data for the synthesis task and achieve excellent results.

REFERENCES

[1] A. Nakano, R. Chan, and J. Hoshino, "Panorama-based immersive story environment," in *Proc. ACM SIGCHI Int. Conf. Adv. Comput. Entertainment Technol.*, Jun. 2004, p. 354.
[2] V. Charissis, "An enquiry into VR interface design for medical training: VR augmented anatomy tutorials for breast cancer," *Proc. SPIE*, vol. 6804, 2008, Art. no. 680404.
[3] T. Cochrane, "Mobile VR in education: From the fringe to the mainstream," *Int. J. Mobile Blended Learn.*, vol. 8, no. 4, pp. 45–61, 2016.
[4] S. Cao, C. Wang, and L. I. Can, *VR/AR Application Innovation and Trend in Cultural Tourism and Film*. Beijing, China: Science & Technology Review, 2018.
[5] E. Upenik and T. Ebrahimi, "A simple method to obtain visual attention data in head mounted virtual reality," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2017, pp. 73–78.
[6] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein, "Saliency in VR: How do people explore virtual environments," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 4, pp. 1633–1642, Apr. 2018.
[7] Y. Rai, J. Gutierrez, and P. Le Callet, "A dataset of head and eye movements for 360 degree images," in *Proc. 8th ACM Multimedia Syst. Conf. (MMSys)*, 2017, pp. 205–210.
[8] A. De Abreu, C. Ozcinar, and A. Smolic, "Look around you: Saliency maps for omnidirectional images in VR applications," in *Proc. 9th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, May/Jun. 2017, p. 1–6.
[9] B. Hu, I. Johnson-Bey, M. Sharma, and E. Niebur, "Head movements during visual exploration of natural images in virtual reality," in *Proc. 51st Annu. Conf. Inf. Sci. Syst. (CISS)*, Mar. 2017, pp. 1–6.
[10] X. Corbillon, F. De Simone, and G. Simon, "360-degree video head movement dataset," in *Proc. 8th ACM Multimedia Syst. Conf. (MMSys)*, 2017, pp. 199–204.
[11] W.-C. Lo, C.-L. Fan, J. Lee, C.-Y. Huang, K.-T. Chen, and C.-H. Hsu, "360°; Video viewing dataset in headmounted virtual reality," in *Proc. 8th ACM Multimedia Syst. Conf. (MMSys)*, 2017, pp. 211–216.
[12] Z. Zhang, Y. Xu, J. Yu, and S. Gao, "Saliency detection in 360 degree videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 504–520.
[13] Y. Xu, Y. Dong, J. Wu, Z. Sun, Z. Shi, J. Yu, and S. Gao, "Gaze prediction in dynamic 360immersive videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 5333–5342.
[14] C. Ozcinar and A. Smolic, "Visual attention in omnidirectional video for virtual reality applications," in *Proc. IEEE 10th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, May/Jun. 2018, pp. 1–6.
[15] E. J. David, J. Gutierrez, A. Coutrot, M. P. Da Silva, and P. L. Callet, "A dataset of head and eye movements for 360° videos," in *Proc. 9th ACM Multimedia Syst. Conf. (MMSys)*, Jun. 2018, pp. 432–437.
[16] H. Hu, Y. Lin, M. Liu, H. Cheng, Y. Chang, and M. Sun, "Deep 360 pilot: Learning a deep agent for piloting through 360° sports videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1396–1405.
[17] L. Xu and M. I. Jordan, "On convergence properties of the em algorithm for Gaussian mixtures,". *Neural Comput.*, vol. 8, no. 1, pp. 129–151, 1996.
[18] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Proc. Comput. Vis. Pattern Recognit.*, 1991, pp. 586–591.
[19] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, vol. 46. Hoboken, NJ, USA: Wiley, 2004.
[20] V. Mnih and G. E. Hinton, "Generating more realistic images using gated MRF," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 2002–2010.
[21] T. Starner and A. Pentland, "Real-time american sign language recognition from video using hidden Markov models," in *Motion-Based Recognition*. Dordrecht, The Netherlands: Springer, 1997, p. 227.
[22] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
[23] R. Salakhutdinov and G. E. Hinton, "Deep Boltzmann machines," in *Proc. AISTATS*, vol. 1, 2009, p. 3.
[24] Z. Tu, "Learning generative models via discriminative approaches," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
[25] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.
[26] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1884–2020.
[27] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," Aug. 2016, *arXiv:1608.06993*. [Online]. Available: https://arxiv.org/abs/1608.06993
[28] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[29] J. Gutièrrez, E. J. David, A. Coutrot, M. P. Da Silva, and P. Le Callet, "Introducing UN Salient360! Benchmark: A platform for evaluating visual attention models for 360° contents," in *Proc. Int. Conf. Qual. Multimedia Exper. (QoMEX)*, Sardinia, Italy, May 2018, pp. 1–3.
[30] K. Regmi and A. Borji, "Cross-view image synthesis using conditional GANs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3501–3510.
[31] M. Abadi, P. Barham, and J. Chen, "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Conf. Operating Syst. Design Implement.*, Savannah, GA, USA, 2016.
[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Dec. 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980
[33] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," Nov. 2016, *arXiv:1611.07004*. [Online]. Available: https://arxiv.org/abs/1611.07004
[34] T. Han, "Alternating back-propagation for generator network," in *Proc. 31st AAAI Conf. Artif. Intell.*, Feb. 2017, pp. 1976–1984.
[35] T. Han, "Divergence triangle for joint training of generator model, energy-based model, and inferential model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8670–8679.

DANDAN ZHU received the Ph.D. degree from Tongji University, Shanghai, China, in 2019. She is currently a Postdoctoral Fellow with the Artificial Intelligence Institute, Shanghai Jiao Tong University. Her research interests include multimedia signal processing, computer vision, and artificial intelligence.

QIANGQIANG ZHOU received the B.S. degree from Jiangxi Normal University, Nanchang, China, in 2003, the M.S. degree in communication and information system from Central South University, Changsha, China, in 2009, and the Ph.D. degree from Tongji University, in 2018. He is currently a Senior Lecturer with the School of Information and Computer, Shanghai Business School, Shanghai, China. His research interests include pattern recognition, machine learning, and computer vision.



TIAN HAN received the B.S. degree in applied mathematics from the Hefei University of Technology, in 2010, the M.Phil. degree in computer science from The Hong Kong University of Science and Technology, in 2013, and the Ph.D. degree in statistics from the University of California at Los Angeles, Los Angeles, in 2019. He is currently an Assistant Professor with the Department of Computer Science, Stevens Institute of Technology. His research interests include statistical machine learning, computer vision, and artificial intelligence.

YONGQING CHEN received the B.S. degree from the South China University of Technology, Guangzhou, China, in 2009, and the M.S. degree in software engineering from Sichuan University, in 2015. He is currently the Chief Engineer with the Hainan Air Traffic Management Sub-Bureau, Haikou, China. His research interests include pattern recognition, machine learning, and computer vision.