Panoramic Image Generation: From 2-D Sketch to Spherical Image

Yiping Duan, *Member, IEEE*, Chaoyi Han, Xiaoming Tao, *Member, IEEE*, Bingrui Geng, Yunfei Du, and Jianhua Lu, *Fellow, IEEE*

***Abstract*—The 360-degree video/image, also called an omnidirectional video/image or panoramic video/image, is very important in some emerging areas such as virtual reality (VR). Therefore, corresponding image generation algorithms are urgently needed. However, existing image generation models mainly focus on 2-D images and do not consider the spherical structures of panoramic images. In this article, we propose a panoramic image generation method based on spherical convolution and generative adversarial networks, called spherical generative adversarial networks (SGANs). We adopt the sketch map as the input, which is a concise geometric structure representation of the panoramic image, e.g., comprising approximately 7% of the pixels for a 583×1163 image. Through adversarial learning, a realistic-looking, plausible and high-fidelity spherical image can be obtained from the sparse sketch map. In particular, we build a dataset of the sketch maps using a visual computation-based sketching model. Then, by optimizing SGANs with GAN loss, feature matching loss and perceptual loss, realistic textures and details are recovered gradually. On one hand, it is an improvement using the sparse sketch map as input rather than the denser input, e.g., the features of the textures and colors. On the other hand, spherical convolution helps to remedy space-varying distortions of the planar projection. We conduct extensive experiments on some public panoramic image datasets and compare them with state-of-the-art techniques to validate the superior performance of the proposed approach.**

***Index Terms*—Panoramic image generation, generative adversarial networks, spherical convolution, sparse sketch map.**

I. Introduction

With the increasing development of multimedia technologies and applications such as smart tourism, virtual reality, and unmanned driving, panoramic images have become a popular topic in recent years with their full-view features. The panoramic image is also called a 360-degree image or omnidirectional image [1]. Recent years have witnessed considerable research efforts in 360-degree image processing. In general, 360-degree image processing is a new multimedia technology for improving the quality of experience (QoE). The panoramic image generation is essential in investigating the QoE of viewing panoramic images. However, the generation model of a panoramic image is significantly different from that of a traditional image. The main difference is that a panoramic image offers an immersive and interactive viewing experience, as the viewers are able to freely move their heads in the range of $360° \times 180°$ to access different viewpoints. Therefore, the generation models of panoramic images and corresponding applications deserve to receive more attention [2].

The image generation problem is essentially an inverse problem. It can be formulated as,

$$\mathbf{M} = f(\mathbf{y}) + \boldsymbol{\xi} \tag{1}$$

where $\mathbf{M}$ is a generated image. $\mathbf{y}$ is a measurement vector. $f(\cdot)$ is the function that recovers the original signal from the measurements. $\boldsymbol{\xi}$ denotes the generation error. From the perspective of information theory, the more measurements we observe, the better the quality of images we can thereby generate [3]. Transform coding [4] is a traditional image reconstruction method. The transform coding method achieves compression of the signal by storing only a large transform coefficient of the signal. When reconstructing the signal, it is only necessary to set the coefficient that is not stored to 0, and then reconstruct the signal through the corresponding inverse transform. Afterward, compressed sensing [5] and sparse representation [6] are employed to perform signal reconstruction at a sampling rate much lower than that of the Nyquist frequency. It is based on dictionary learning and sparse representation theory, using sparse coefficients for signal reconstruction.

Recently, GANs [7] have been shown to be capable of generating visually satisfactory images, making it difficult even for humans to determine the authenticity of generated images. In addition, GANs generate images with a relatively small amount of data, such as a noise vector, or semantic label map [8]. Therefore, GANs have achieved promising results in natural image generation on various datasets, such as the VGG face dataset [9], and the ImageNet dataset. However, the panoramic images are fundamentally different from natural images due to the different imaging mechanism. Specifically, the panoramic

Manuscript received April 21, 2019; revised September 6, 2019; accepted November 19, 2019. Date of publication January 22, 2020; date of current version February 5, 2020. This work was supported in part by the National Key R&D Program of China under Grant 2018YFF0301205, in part by the National Natural Science Foundation of China NSFC under Grants 61801260 and 61925105, in part by the Chinese Postdoctoral Science Foundation under Grant 2018T110098, and in part by iQIYI. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Ali Borji. *(Corresponding author: Xiaoming Tao.)*

Y. Duan, C. Han, X. Tao, B. Geng, and J. Lu are with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China, and with the Beijing National Research Center for Information Science and Technology, and also with the Beijing Innovation Center for Future Chip, Beijing 100084, China (e-mail: yipingduan@mail.tsinghua.edu.cn; hancy16@mails.tsinghua.edu.cn; taoxm@mail.tsinghua.edu.cn; gengbr@126.com; lhh-dee@mail.tsinghua.edu.cn).

Y. Du is with the Department of Electrical and Information Engineering, and also with the State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China (e-mail: 531187226@qq.com).

Digital Object Identifier 10.1109/JSTSP.2020.2968772

Fig. 1. A spherical panoramic image is projected onto the 2-D plane. Left is the spherical image, and right is the plane image. It is observed that some distortions exist in the plane image, such as the road and the tree.

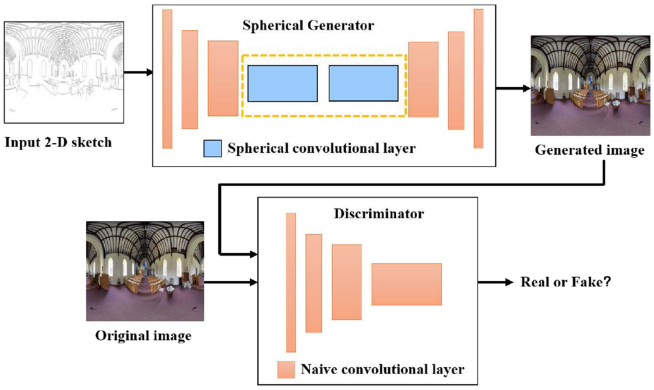


Fig. 2. The framework of the proposed approach with the paradigm of SGANs. Using the 2-D sketch map as input, the spherical generator is used to synthesize the panoramic image. The discriminator is used to distinguish the generated images from the real images.

image consists of spherical images captured at different viewpoints. When using the naive convolution, the spherical images need to be projected to a planar space. This projection will produce space-varying distortions, as shown in Fig. 1. Noted that some distortions exist in the plane image, such as the road and tree. On the other hand, most existing works do not consider the structures of images during generation (when generates from random noises or semantic maps), which are important to users' QoE. Therefore, it is necessary to develop specific generation methods for panoramic images.

In this paper, we propose a panoramic image generation method to automatically translate a 2-D sketch to its spherical form. In fact, it is a way to balance the conflict between the input data amount and the high quality. In particular, we build a dataset of sketch maps using a visual computation-based sketching model. Then, through optimizing SGANs with GAN loss, feature matching loss and perceptual loss, the textures and details are recovered gradually, and finally a spherical panoramic image can be generated. The whole network is trained end-to-end from the sparse contour to the spherical image. The framework of the proposed approach is illustrated in Fig. 2. With the contour of a panoramic image, our model can generate the texture and details using the statistical prior from the large dataset. When viewed as a communication system, this will bring in a great reduction on the volume of transmitted data (from panoramic image to sketch map), as shown in Fig. 3. For a panoramic image of size 583×1163, the sketch map accounts for approximately 7% of the total

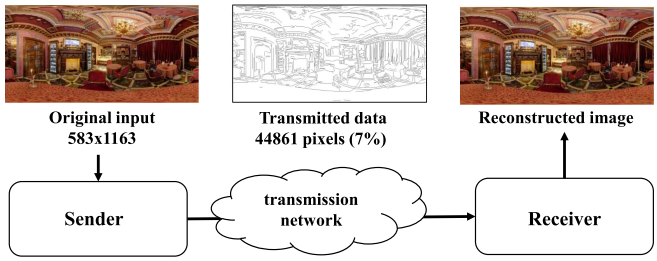


Fig. 3. A communication system using the sketch representation. For the sender, it extracts the sketch map of the original signal/image. Sketch points only account for seven percent of the image pixels. For the receiver, it generates the reconstructed image using the transmitted sketch map. It will greatly reduce the amount of transmitted data.

pixels with only 44,861 sketch points. The main contributions of this paper are two-fold:

- We propose a generative adversarial network with spherical convolution for panoramic image generation, which reduces the distortion caused by 2-D planar mapping.
- We develop the contour-based generation method. Using the sparse sketch map as the input, the textures and details can be filled and the spherical image is generated by adversarial learning. It is a significant improvement from the 2-D sparse sketch map to the spherical image.

This paper is organized as follows. Section II reviews the related work of image generation. Section III describes how we construct the sparse sketch map of the panoramic image. Simultaneously, we propose the panoramic image generation model based on SGANs. Section IV reports the experimental results, and Section VI concludes this paper.

II. Related Work

With the increasing demands for image quality in visual applications, image generation has been an ongoing hot topic in the field of computer vision and image processing. Previous panoramic image generation methods mainly focus on the stereo imaging system, such as the sensor, input-output devices and so on [10]. Our work mainly treat the panoramic image generation problem from the viewpoint of signal processing. Here we briefly review the relevant image reconstruction/generation literatures from three aspects: transform coding, compression sensing and deep learning.

A. Transform Coding-Based Reconstruction/Generation

Linear transformation of the image operates by transforming the image from the spatial domain to another transform domain. After linear transformation, the image compression can be realized by finding the appropriate quantization and entropy coding methods according to the rate distortion theory. As long as the distortion of the reconstructed image is within the allowable range, any compression ratio can be used with greater flexibility. Therefore, in the previous research of image processing, image reconstruction based on transform coding was the most common algorithm. In [11], the authors examined a set of orthogonal moment functions based on the discrete Tchebichef polynomials, and the effectiveness of Tchebichef moments as feature

descriptors was proved in image reconstruction. Elshoura *et al.* [12] proved that Tchebichef moment was better than Legendre and Zernike moments in image reconstruction regardless of the nature of the image used by analyzing image reconstruction accuracy in the frequency domain. In [13], Hong *et al.* proposed an adaptive reconstruction method for loss blocks of the image. The discrete cosine transform (DCT) coefficients were recovered by adaptively selecting available neighboring blocks and interpolation. The experimental results showed that the image quality was restored successfully and the computational complexity is low. Gunturk *et al.* [14] proposed a random framework to efficiently use quantitative operations, transform-domain statistics and additive sensor noises for image superresolution reconstruction. In addition, the wavelet transform is a time-frequency analysis method. It has good positioning performance in the time domain and frequency domain. For a given continuous signal $f(t) \in L^2(R)$, the wavelet transform can be expressed as,

$$f(t) = \sum_k \beta_{j_0,k}(t)\varphi_{j_0,k}(t) + \sum_{j=j_0}^{\infty}\sum_k \alpha_{j,k}\Psi_{j,k}(t) \quad (2)$$

where $\varphi_{j_0,k}(t)$ is the scaling function and $\Psi_{j,k}(t)$ is the wavelet function. The emergence of the wavelet transform has made image processing more convenient [15] and has become one of the classic methods of image reconstruction. In [16], Starck *et al.* performed multi-resolution deconvolutions on the interference image using wavelet coefficients and reconstructed the images of two evolutionary stars from the infrared speckle interferometry. The discrete wavelet transform was also applied to synthetic aperture radar (SAR) image generation problems [17].

B. Compressed Sensing-Based Reconstruction/Generation

Different from the above transform coding methods, a number of compressed sensing-based approaches for image reconstruction algorithms are proposed. Compressed sensing [5] can sense signals at a rate much lower than that of the Nyquist sampling theorem, while retaining the basic information contained in the signal. Image reconstruction based on compressed sensing estimates the original image reconstruction from the image compression observation, which usually requires complex calculation to solve the highly nonlinear optimization reconstruction problem. The goal of compressed sensing is to recover the signal z from its measurements y. The basic model of compressed sensing image reconstruction is written as follows:

$$\mathrm{z}^* = \arg\min_{\mathrm{z}} \|\mathrm{z}\|_0 \quad \text{s.t. } \|\mathrm{y}-\Phi\mathrm{z}\|^2 \le \varepsilon \quad (3)$$

where $\|\cdot\|_0$ is called the l_0 norm and used to calculate the number of non-zero elements in the signal. Φ is a sensing matrix. y is a measurement vector. It is a non-convex optimization term that measures the sparsity of the signal. When the sparsity of the signal is a known condition, the reconstruction can also be perfromed by solving the following model.

$$\mathrm{z}^* = \arg\min_{\mathrm{z}} \|\mathrm{y}-\Phi\mathrm{z}\|^2, \ \text{ s.t. } \|\mathrm{z}\|_0 \le K \quad (4)$$

where K is the sparsity of the signal.

The optimization problem of reconstructing coefficient signals from compressed sensing is an NP-hard problem [18]. In addition to sparsity, fully exploiting other prior knowledge has produced rich image reconstruction strategies and methods, mainly including greedy algorithms [19], [20] and convex relaxation algorithms [21]–[23]. The sparse decomposition of signals under a redundant dictionary was the most classic greedy iterative algorithm proposed by Mallat *et al.* [24]. Because of the small amount of calculation and easy implementation, it has been widely used in image compressed sensing reconstruction [25], [26]. Subsequently, many scholars improved the above method, and proposed an orthogonal matching pursuit (OMP) algorithm [27]–[29], which orthogonalized the column vectors in the selected sub-matrix, so that the algorithm converged faster. In [30], the authors proposed a regularized orthogonal matching pursuit degree algorithm which was more suitable for image signal reconstruction. Thong *et al.* introduced the sparsity adaptive matching pursuit algorithm in [31]. Another common algorithm is to transform the non-convex optimization problem in the compressed sensing reconstruction into a convex optimization problem. In [32], the authors proposed an iterative shrinkage threshold method, which was a commonly used convex optimization algorithm for compressed sensing reconstruction. [33] introduced the Bregman iterative regularization method into compressed sensing reconstruction and transformed the compressed sensing reconstruction problem into an unconstrained optimization problem. A method to perform image reconstruction using highly incomplete data by a special recursive filtering process was proposed [34]. In [35], an image reconstruction method based on an autoregressive model is proposed. Yu *et al.* used a mixed Gaussian model for training and reconstruction [36]. These methods mined and utilized the statistical prior information of the image in the image sparse representation and reconstruction model and were able to obtain fast recovery and estimation of the image. The image reconstruction algorithm based on non-local central clustering proposed by Dong *et al.* was also a classical reconstruction algorithm based on sparse characteristics [37], [38].

C. Deep Learning-Based Reconstruction/Generation

In recent years, the rapid development of deep learning [39]–[41] has brought new ideas and perspectives to the method of image reconstruction/generation. In the work [42], Lohit *et al.* proposed a data-driven noniterative algorithm to overcome the shortcomings of early iterative algorithms and used end-to-end deep convolutional neural networks (CNNs) to learn the measurement matrix and reconstruction algorithm in a single network. Moreover, using CNNs, the authors extracted a hierarchy of increasingly spatial features for reconstructing hyperspectral imagery, and the results were better than traditional methods [43]. In addition, in the work [44], the authors provided a new method for holographic image reconstruction based on deep learning. This algorithm can quickly eliminate twin-image and self-interference-related artifacts using only one hologram intensity. The neural network framework was more efficient than existing holographic phase recovery methods. Kelly *et al.* [45]

used CNNs as a quasi-projection operator in the process of least squares minimization to encode high-level information in the image. The proposed method improved the performance of image reconstruction in iterations. GANs [7] model the distribution of natural images by forcing samples that are indistinguishable from the original images. This algorithm is widely used in image generation. Radford *et al.* [46] proposed a more stable GAN; this network learnt good representations of images for supervised learning and generative modeling. Reed *et al.* proposed a new model, named generative adversarial what-where network [46]. The images generated through this network can give instructions about what content to draw in which location. In the work [47], a new GAN algorithm, which can improve learning stability and eliminate mode collapse, was proposed and named Wasserstein GANs. The energy-based generative adversarial network model was proposed and trained to generate high-resolution images in the work [48]. Isola *et al.* [49] investigated conditional adversarial networks for image-to-image translation. This method was effective at synthesizing images from label maps. Afterward, Zhu *et al.* [8] proposed an approach for learning to translate an image from a source domain to a target domain in the absence of the paired examples. Wang *et al.* [50] proposed multi-scale generator and discriminator architectures for high-resolution image generation.

III. Proposed Model

A. 2-D Sketch

The concept of "sketch" dates back to the seminal work of David Marr [51]. At the time, Marr did not provide a strict mathematical proof. Afterward, Songchun Zhu *et al.* [52] continued the work of Marr and provided a mathematical model of the sketch. A sketch map is a concise and structural representation of the image. It represents the positions of the amplitude or intensity changes. In addition, it also captures the shape and geometric structures of the image. In Zhu's paper, they divided the images into the structure domain and texture domain according to the primal sketch. Then, a sparse coding model was used to represent image intensities of the structure domain, such as edges and ridges. The Markov random field model was used to summarize the intensities of the texture domain with some regular pattern. In [53], the ratio of verage detector (ROA) and cross-correlation-based (CC) detector were fused to detect the edge-line features of the SAR images. These two detectors are also equally applicable to natural images and panoramic images. We use the sketching model in [53] to compute the sketch map of the panoramic image. Simultaneously, the orientations of the edges are also computed in the sketching process. There are three parameters in the model of [53], including the coding length gain (CLG), the high threshold (HT) and low threshold (LT) of the non-maximum suppression (NMS). The larger the CLG is, the sparser the sketch map is, i.e., the fewer the number of segments in the sketch map. It should be noted that the sketch map is obviously different from traditional edge detectors, such as the Canny detector. In fact, the filter of the Canny detector can be used as a subset of the filters in the sketching process. In addition, the contour map of the Canny detector consists of edge points. The sketch map consists of line segments with the orientation information and can be derived from the following steps [53].

1) Compute the edge-line intensity map. The edge-line intensity map is calculated by adaptively fusing the responses obtained by edge-line templates with different scales and orientations.
2) Extract the sketch curves. Based on the intensity map, the non-max suppression and double-threshold based connection methods are used to extract the curves contained in the images.
3) Extract the sketch lines. By the approximation method, the extracted curve is represented as a sketch line.
4) Generate the sketch map. The significance of each sketch line is evaluated by a pair of paradoxical hypotheses. Then, the significant sketch lines are preserved to constitute the sketch map.

Fig. 4 shows the sketch maps of some panoramic images. Fig. 4(a) is the original panoramic images. Fig. 4(b) is the 2-D sketch maps of the panoramic images with the threshold $CLG = 15, HT = 1.5$ and $LT = 0.7$. Fig. 4(c) is the edge maps of the Canny detector with $\sigma = 2$. Fig. 4(d) is the edge maps of the Canny detector with $\sigma = 4$. We can see that the sketch maps are clearer than the edge maps of the Canny detectors. For example, in the second row, the sketch map represents the structures of the buildings. However, the Canny detector just describes the positions of the edges. Moreover, the sketch line segments or the edge points always account for a small percentage of the total image pixels and the corresponding percentages are shown in Fig. 4. Analogous to [50], [54], the sketch maps also use the boundary map as the input. The difference is that sketch maps not only contain the boundary but also the shape and geometric structures. The code of the sketching model is available at http://web.xidian.edu.cn/fliu/lunwen.html

The reasons we select the sketch map as inputs are described as follows. First, the sketch map is a concise structure representation of the image. It contains more details and is helpful for improving the generated quality. Nevertheless, the semantic map only contains the label information. In other words, the sketch map contains more structures than the semantic map. Certainly, both the sketch map and the semantic map can be used as the input to generate the panoramic image. Second, we usually need complex computations to produce the semantic map in an supervised manner. However, the sketch map is computed by the predefined detectors in a unsupervised manner. Third, in the communication system, if the transmitted data are represented as a sketch map, it will greatly reduce the transmitted data volume of the communication. For example, the panoramic image of size 583×1163 has a sketch map that contains 44,861 sketch points, accounting for approximately 7% of the total pixels.

B. Spherical Convolution

In recent years, CNNs have achieved breakthrough results in speech recognition, visual object recognition, natural language processing, and other fields. CNNs are very good at analyzing signals such as audios, texts, images, or videos. Their shift-invariance provides it a powerful discriminating power in

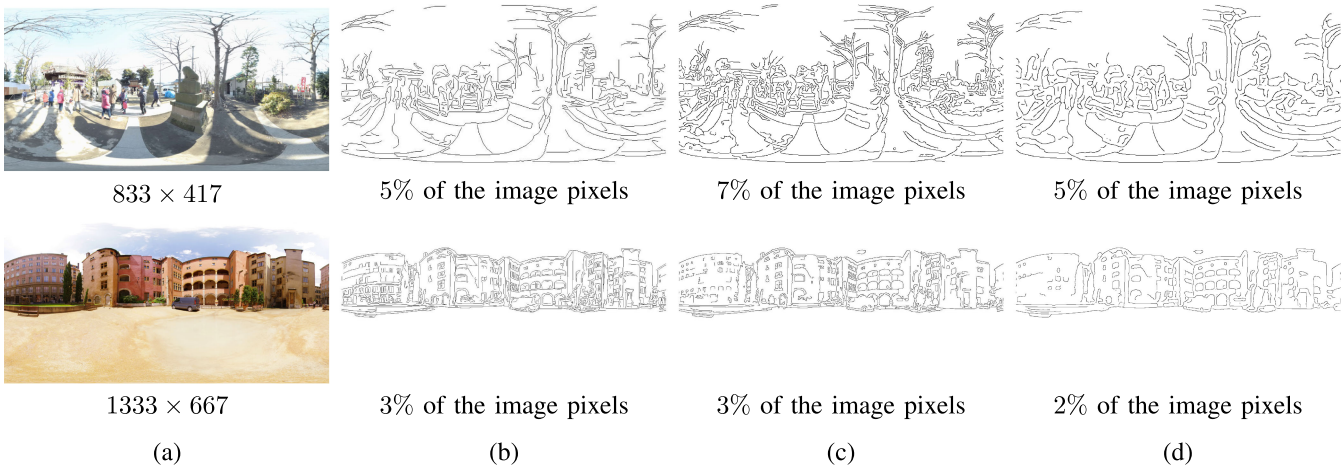


Fig. 4. Sketch maps of panoramic images. (a) is the original panoramic images. (b) is the corresponding sketch maps of the original images. (c) is the edge contour using the Canny detector with $\sigma = 2$. (d) is the edge contour using the Canny detector with $\sigma = 4$.

computer vision. However, the current CNN models do not have rotational invariance and cannot observe data from multiple angles during the training process. However, many of today's applications, such as virtual reality, autopilot, etc., use a 360-degree full camera to capture the spherical signals. Currently, spherical signals are usually mapped to a two-dimensional space and processed using naive CNNs. These projections can cause shape and size distortions. Moreover, some areas look larger or smaller than the actual scene.

In [55], the authors pointed out that the equivariance is a good inductive bias in CNNs and proposed group equivariant CNNs. Afterward, they naturally extended the two-dimensional system to the three-dimensional system and proposed a spherical CNN [56]. It can be used in the recognition of 3-D shapes and many prediction tasks. In particular, $S2$ is a two-dimension manifold in the spherical space and defined as the set of points $x \in \mathbb{R}^3$ with a norm of 1. The output feature map is computed as the inner product between the input feature map and the filter that has been rotated by R. According to [56], for spherical signals u and filter ψ, spherical correlation is defined as,

$$[\psi * u](R) = \langle L_{R\psi}, u \rangle = \int_{s^2} \sum_{k=1}^{K} \psi_k(R^{-1}x) u_k(x) dx \quad (5)$$

where $\langle \cdot \rangle$ is the inner product. dx is the standard rotation-invariant integration measure on the sphere.

The rotation group $SO(3)$ is a three-dimensional manifold and defined as the set of rotations. Similar to $S2$, the rotation group correlation is defined as,

$$[\psi * u](R) = \langle L_{R\psi}, u \rangle = \int_{sO(3)} \sum_{k=1}^{K} \psi_k(R^{-1}Q) u_k(Q) dQ \quad (6)$$

where dQ is the invariant integration measure on $SO(3)$.

To solve the above equation, Cohen *et al.* [56] used the fast Fourier Transform (FFT) to compute the correlations. The complexity of the spherical convolution is dependent on the FFT and represented as $O(n \log n)$. The spherical convolution can be computed by the following steps,

- Compute the FFT coefficients on the spherical signals u and the filters ψ, separately. The two coefficients are defined as $\widehat{u}$ and $\widehat{\psi}$;
- Compute the product of FFT coefficients $\widehat{u}$ and $\widehat{\psi}$;
- Compute the inverse FFT on the product to obtain the output feature map.

C. Panoramic Image Generation Model

Panoramic images always have high resolutions, which introduce great challenges for their generation. Previously, DNN-based methods mainly focused on thumbnail images, such as the MNIST dataset with the size of 28×28, the CIFAR dataset with the size of 32×32 [57] and so on. Recently, some studies began to generate the high-resolution images with DNN-based method, such as [8], [50]. In the proposed approach, we mainly improve the pix2pixHD framework [50] by using the spherical convolution and 2-D sketch. Under the GAN framework, the proposed approach includes the generator G and the discriminator D. For our task, the generator is used to translate the 2-D sketch into realistic-looking panoramic images. The discriminator is used to distinguish the generated images from the real ones. After the adversarial learning, a sketch map is translated into a spherical image.

Generator: The aim of the generator is to synthesize the spherical panoramic image from the 2-D sketch. At the same time, we hope that the generated image is close enough to the original image. We design the spherical generator network to generate the spherical image. By using the spherical convolution, it can reduce the distortion due to the planar space projection. The training data comprise a set of pairs of images $\{(s_i, x_i)\}$, where s_i is the 2-D sketch and x_i is the original panoramic image. To improve the fidelity and perceptual quality of the panoramic image, the generator is designed from three aspects. First, the generator G generates the image, which is hard for the discriminator D to distinguish from the real image. We use the

loss of least squares GANs (LSGANs) to describe this constraint,

$$\mathcal{L}_{GAN} = \mathbb{E}_{s\sim p_{\text{data}}(s)}[(D(s, G(s)) - 1)^2] \tag{7}$$

Second, we use feature matching loss to make the training process stable. Analogous to [50], for the generated images and the real images, we extract the features from the discriminator at different layers. On one hand, the feature makes the training process stable. On the other hand, it is a kind of perceptual loss to improve the quality of the generated image. Assume that L_1 is the total number of the layers. F_D^i represents the feature extracted from the ith layer from the discriminator. The feature matching loss is expressed as,

$$\mathcal{L}_{FM} = \mathbb{E}_{(s,x)} \sum_{i=1}^{L_1} \frac{1}{N_i}[\|F_D^i(s, x) - F_D^i(s, G(s))\|_1] \tag{8}$$

where N_i is the number of the elements in the ith layer.

Third, the perceptual quality is an important evaluation index in image generation, especially using the DNN-based methods. Achieving plausible QoE to the generated image is the ultimate goal. Therefore, we should emphasize this ability when constructing the model. Through optimizing the similarity in the feature space, it can improve the semantic similarity between the generated image and the real image. This loss has been used in several works to improve the perceptual quality [50], [58], [59]. In particular, for the generated image $G(s)$ and the real image x, their features are extracted from the VGG network. The corresponding perceptual loss is computed as,

$$\mathcal{L}_{\text{perceptual}} = \sum_{i}^{L_2} \frac{1}{M_i}[\|F_{VGG}^i(s, x) - F_{VGG}^i(s, G(s))\|_1] \tag{9}$$

where L_2 is the total number of layers in the VGG network. M_i is the number of elements in the ith layer. F_{VGG}^i is the feature extracted from the ith layer of the VGG network.

Based on the above, we combine the three losses as the objective function of the generator,

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{GAN} + \mathcal{L}_{FM} + \mathcal{L}_{perceptual} \tag{10}$$

Discriminator: The aim of the discriminator is to distinguish the generated images and the real images. For real images, the discriminator wants its output to be true. For generated images $G(s)$, the discriminator wants its output to be false. The loss of the discriminator is written as,

$$\mathcal{L}_G = \frac{1}{2}\mathbb{E}_{(s,x)\sim p_{data}(s,x)}[(D(s, x) - 1)^2] + \frac{1}{2}\mathbb{E}_{s\sim p_{data}(s)}[(D(s, G(s)))^2] \tag{11}$$

In addition, we use multi-scale discriminators D_1, D_2 and D_3, which is common in image generation [50]. The coarse to fine paradigm can improve the quality of the generated image. At the coarse scale, it can capture the global information and improve the consistency of the generated image with the large receptive field. At the fine scale, it captures the information at the local view and preserves the details, such edges and lines. Moreover, the multi-scale methods can reduce the burden on the network and make the network easier for training.

Algorithm 1: Panoramic Image Generation Algorithm.

Initialize: Resize the input to 512 × 512 px, set the batch size to 1, the learning rate of the deep neural network as 0.0002, and the maximum iteration number of training the network as 100 epochs. Assume that (s_i, x_i) is a pair, s_i is the 2-D sketch map, x_i is the real image, G is the generator, and D is the discriminator.

1: **while** current epoch $\leq epoch\_max$, **do**
2: Feed s_i into the spherical generator G to generate the fake image $G(s_i)$;
3: Concatenate the sketch map and the fake image as $input_D1 = [s_i, G(s_i)]$;
4: Feed $input_D1$ into the discriminator to produce $D(input_D1)$; then, using the loss of LSGANs to obtain $Loss\_D\_fake$ and $Loss\_G\_GAN$;
5: Concatenate the sketch map and the real image as $input_D2 = [s_i, x_i]$; feed $input_D2$ into the discriminator to produce $D(input_D2)$; then, using the loss of LSGANs to obtain $Loss\_D\_real$;
6: Extract the features from the discriminator; $F_D^i(s, x)$ is the extracted features of the real image; $F_D^i(s, D(input_D1))$ is the extracted features of the fake image; then, using the L_1 loss to obtain $Loss\_FM$;
7: Extract the features from the trained VGG network; $F_{VGG}^i(s, x)$ is the extracted features of the real image; $F_{VGG}^i(s, D(input_D1))$ is the extracted features of the fake image; then, using the L_1 loss to obtain $Loss\_VGG$;
8: The loss of the discriminator is, $L_D = Loss\_D\_fake + Loss\_D\_real$, The loss of the generator is, $L_G = Loss\_G\_GAN + Loss\_FM + Loss\_VGG$,
9: Fix the network parameters of the discriminator, optimize and update the generator using (10);
10: Fix the network parameters of the generator, optimize and update the discriminator using (11);
11: **end while**

D. Network Architecture

We improve the network architecture in [50] and maintain the same settings for common hyperparameters. The network architecture of the proposed approach is shown in Fig. 5. In fact, it is a forward process of the data flow. From left to right, the input is the 2-D sketch map. Through the spherical generator network, discriminator network and VGG network, five losses are computed. In the backward process, we optimize the losses to train the model.

The generator adopts the "Encoder-Decoder" structure. The input of the generator is the 2-D sketch, which needs to be projected onto the sphere by the method used in [56] for preprocessing. We follow the definition used in [56] and [50]. Assume that spherical convolution $S2$, normalization and activation function is a tuple and defined as $S2 - k - b$, where

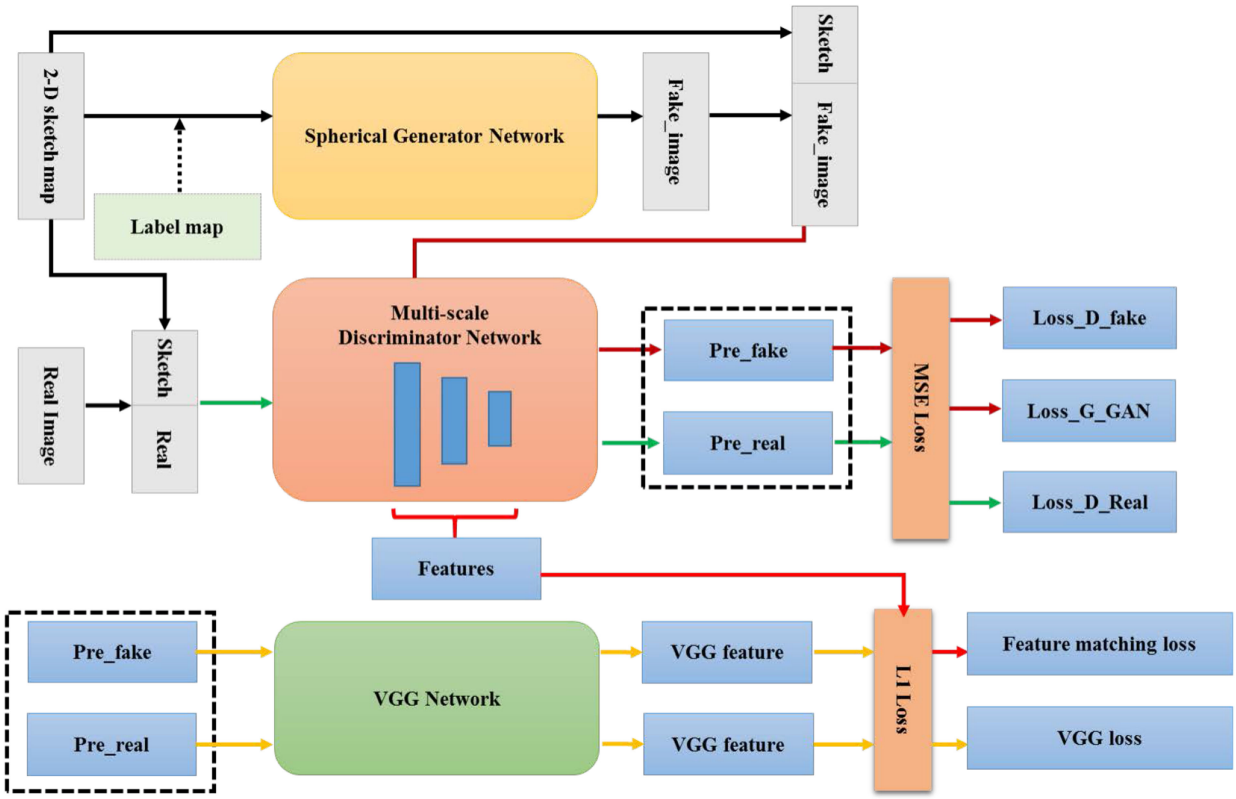


Fig. 5. The structure of the proposed model. It shows the flow of data from input to output and the froward process of the proposed model. Through the generator and the discriminator, we will obtain five losses. By optimizing the network through those loss functions, we will obtain the expected model.

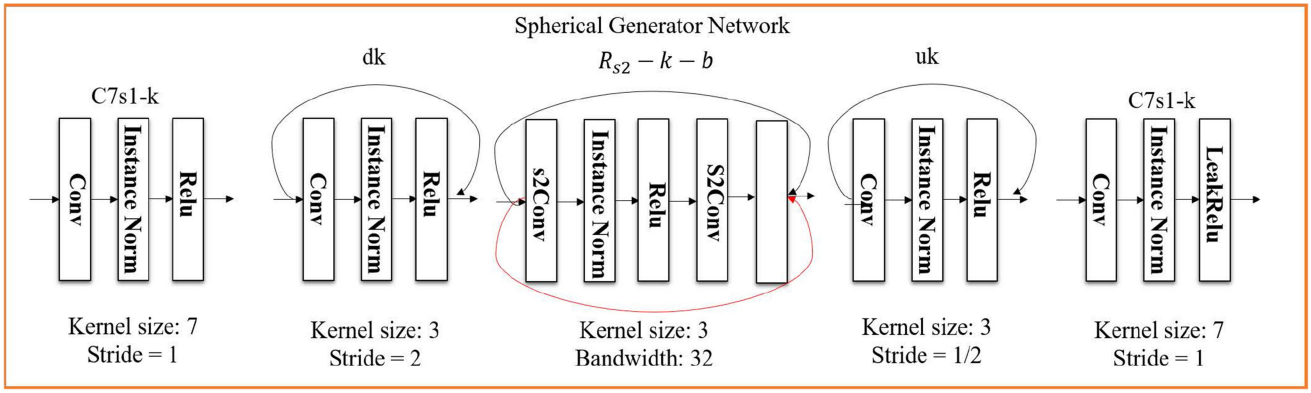


Fig. 6. The structure of the generator. The output of every $S2$ convolution is averaged along the γ dimension to transform the feature map to 2-D again. The black arrow represents that there are several corresponding tuples in the network. The red arrow represents the residual block.

k is the number of the input/output channel and b is the input/output bandwidth. R_{S2-k-b} is a residual block that consists of two $S2-k-b$ tuples. Here, we choose ReLU as the activation function and Instance3d as normalization. The generator started with a naive convolutional layer with the kernel size of 7×7. The second block is composed of three downsampling layers. Then, 5 residual blocks follow. Symmetrically, three upsampling layers are included. Finally, a convolutional layer is used to produce the 3-dimensional output. Using similar notations to [50], the spherical generator network contains: $c7s1-26, d208, d104, d52, (R_{S2-208-32}) \times 5, u52, u104, u208$ and $c7s1-3$. Here, $R_{S2-208-32}$ denotes $S2$ convolution with the 3×3 kernel, input/output channel 208, and input/output bandwidth 32. The generator structure is shown in Fig. 6. We do not use $SO(3)$ convolution because it requires a huge amount of memory and computation. Moreover, the rotation along the γ dimension is not included in our experiments (dataset). The $S2$ convolution, to some extent, results in less memory usage.

For the discriminator, we follow the same network architecture as [50]. We have not used spherical convolution in the discriminator for two reasons. First, the main aim of the

Fig. 7. Some real images from AOI dataset.

discriminator is used to distinguish the generated images from the real images. This aim is not closely related to whether the image is two-dimensional or spherical. Second, using spherical convolution will increase the storage and computational burden of the network. Therefore, we only use spherical convolution in the generator to generate the spherical image. It avoids distortion of the image and improves the generated quality of the image.

IV. Result and Analysis

In this section, we conduct experiments with two parts to verify the effectiveness of the proposed approach. In the first part, we only use the sketch map as input to synthesize the panoramic image. The experiments are mainly used to prove that textures can be synthesized where no another input information is provided. In the second part, by introducing the semantic label map, the synthesize quality is improved. The experiments confirm that when more information is transmitted, better quality is achieved. In addition, it should be noted that we train different models for different datasets because the distributions of the two datasets are different. If the two datasets are trained with one model, they may interfere with each other. The steps of the proposed appraoch are shown in Algorithm 1. Our code is available at https://github.com/hancy16/SGAN

A. Results on AOI Dataset

We develop the original panoramic image dataset (AOI dataset) [60] for panoramic image generation. It is available at https://drive.google.com/open?id=1iXZSFXUzI3QK9wBfrJfN3yjIasUiu67m. We construct its corresponding sketch map dataset using the sketching model in Section III. A. The original dataset and its sketch map form a pair as a new dataset for panoramic image generation. This dataset includes 556 panoramic images and 556 sketch maps. These images are collected from different scenarios, classified into Cityscapes, Hall_exhibition, Human_party, Human_tour, Indoor_rooms and Natural_landscapes. The resolutions of the images vary from 299×597 to 1333×667. The sketch map is a binary image. We define the sketch lines as 0 and the rest as 255. However, other values can also be defined, such as 0 and 1. It will not affect the training of the proposed model. We divide the dataset into a training dataset and a testing dataset, where the training dataset includes 500 pairs and the testing dataset includes 56 pairs. To our best knowledge, this is the first dataset using a sketch map for panoramic image generation. Some examples of the AOI dataset are shown in Fig. 7. The AOI dataset and its sketch map dataset are available at https://drive.google.com/drive/u/0/folders/1JjgHm_5HROoeuHA67RYD–GJR8iaFmko

Experimental setup: With the dataset, we train the model on the training dataset from scratch, which contains 500 panoramic images and sketch maps. We use the testing set for testing, which contains 56 pairs. Due to the huge memory burden produced by spherical convolution, the inputs are resized to 512×512 px. All of the convolutional kernels are initialized with a Gaussian distribution with a mean of 0 and a standard deviation of 0.02. The weights of the InstanceNorm3d are initialized with a Gaussian distribution with a mean of 1 and a standard deviation of 0.02. We choose the Adam solver as the optimizer with a learning rate of 0.0002. The batch size is set to 1. We train the proposed model on 4 TITAN XP GPUs. The training is finished after 100 epochs.

Result: The generation results on these panoramic images using only the sparse sketch information are shown in Fig. 8. We resize the generated images to the original size for visualization. In the AOI dataset, the ratio of the sketch lines to the total pixels is approximately 7%. Using the very sparse sketch map, the textures and details are recovered step by step by adversarial learning. In [8], [41], [54], they also use the sparse contours to generate the image. However, their databases always contain a single type of image, such as VGG faces, Stanford Dogs and so on. In addition, the content of the image scene is simple. For different datasets, they train different models. Panoramic images with high resolution always contain rich textures and details, which give rise to great challenges for panoramic image generation. From Fig. 8, we can see that the proposed approach produces high quality reconstruction. In particular, the edge portion of the synthesized image are clear. For example, in the third row of Fig. 8, this panoramic image contains complex scenery, buildings and some people. The proposed model

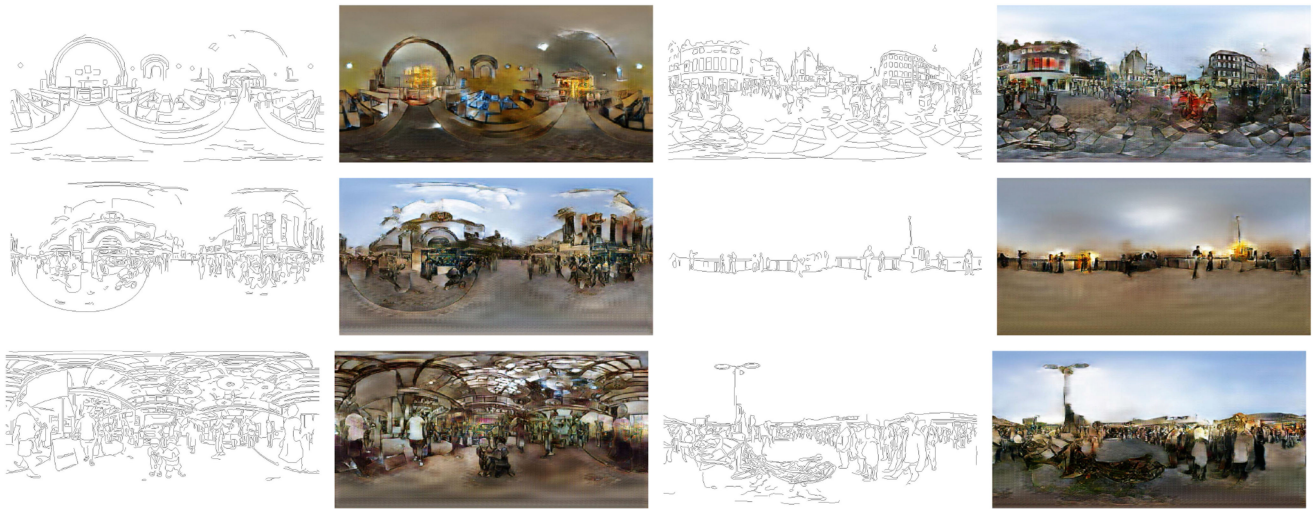


Fig. 8. Visual example of the images generated on AOI dataset. The first and the third column are the sketch maps. The second and fourth column are the synthesized images by the proposed approach.

generates rich and complicated textures and details. From the visual quality, the generated images are plausible and have a realistic-looking appearance. These images can also be used in subsequent processing, such as semantic segmentation, object recognition and other visual tasks.

Comparison: We compare the proposed method with the baseline Pix2pixHD model [50]. The corresponding experimental results are shown in Fig. 9. For a fair comparison, we use the same settings and details for the baseline model and retrain the model. Regarding the Pix2pixHD model, it is an improved version of the Pix2pix model [49]. Using the coarse-to-fine generator, multi-scale discriminator and feature matching loss, the Pix2pixHD model improves the generated quality and is suitable for the generation of high-resolution images. From the third column in Fig. 9, it is observed that the color is strange. For example, in the first image, the colors of the sky and ground are mixed. Moreover, the generated structures of the image are not plausible, especially the fourth image. By using the proposed approach, the visual quality is improved and the textures are recovered clearly. In the proposed method, we plug the spherical convolution into the generator to reduce the distortions. With the proposed model, we implement the panoramic image generation using only the sparse sketch maps. From Fig. 9, we can see that the visual quality of the proposed method is better than that of the baseline model. It illustrates that spherical convolution is necessary for a spherical signal/image. Meanwhile, the perceptual loss is useful in high-level image generation.

B. Results on SYNTHIA Dataset

SYNTHIA dataset: The SYNTHIA dataset [61] is generated from a visual world of urban scenarios with semantic labels available. Such class annotations are shown to be essential for high-quality image generation and, on the other hand, take very little to store, e.g., 0.036 bpp on average for downscaled Cityscapes [62] images as reported in [63]. Moreover, the dataset provides multiview images, which makes it a perfect candidate of the semantic panoramic dataset. Therefore, we conduct the panoramic image generation task on the SYNTHIA dataset and include its semantic label maps.

Spherical projection: The SYNTHIA dataset contains several video sequences with 8 views in each frame. The original data comprises normal images, so we generate panoramic images by projecting them on a sphere. Specifically, for each frame we use the 4 images captured by the left stereo-camera and transform them into a single panoramic image. To complete this task, we first select a reference spherical coordinate system R_0 and suppose that the distances between images and the origin are constant, namely, f. For one image I, we obtain its coordinates in R_0 using the rotation transformation,

$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{bmatrix} \times \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{bmatrix} \times \begin{bmatrix} x - \frac{W}{2} \\ y - \frac{H}{2} \\ f \end{bmatrix} \tag{12}$$

where (x, y) denotes the index in I, (W, H) denotes the width and height of I, (α, β) denotes the rotation relative to R_0, and f is actually the focus of the camera. In our implementation, f is given by $W/[2\tan(\frac{\tau}{2})]$, where τ is the field angle of the camera. We project the images on a sphere with radius f. Then the spherical coordinates can be given as,

$$\theta = \arctan\left(\frac{u}{w}\right)$$
$$\phi = \arctan\left(\frac{v}{\sqrt{u^2 + w^2}}\right) \tag{13}$$

After projecting all images, we average the intensity for overlapped regions and crop the 2-D sphere since the top and bottom views are lacking. The whole process is illustrated in Fig. 10.

Fig. 9. Visual examples of images generated by our approach and baseline model on AOI dataset. From left to right: input sketch map, our approach, pix2pixHD baseline model.

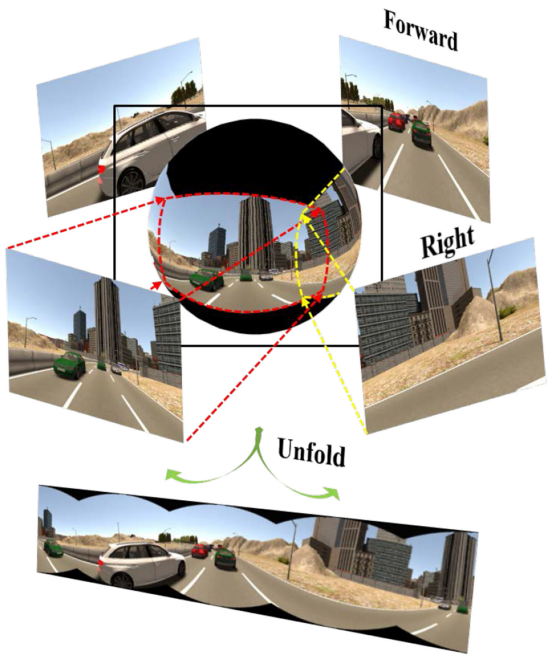


Fig. 10. Illustration of the spherical projection, the 4 views are seamlessly stitched.

TABLE I
PARAMETER SETTING FOR SPHERICAL PROJECTION

View	α	β	τ
Front	0	0	100
Left	0	90	100
Back	0	180	100
Right	0	270	100

The parameter setting is shown in Table I. We downsample the original images with factor 0.5 during processing, and the final panoramic images are all of size 1687 $\times$ 335 px. Some examples of the dataset are shown in Fig. 11. The datasest is available at https://drive.google.com/drive/u/0/folders/10BF1hZQnndj-Yi-2vX7gGwww1f6i8jP9

Experimental setup: We adopt the SYNTHIA-SEQ5-05-SUMMER dataset and generate 787 panoramic images in total. Then 579 images are randomly selected as the training set and the remaining comprise the test set. During training, all images are resized to 512 $\times$ 512 px. We train the networks from scratch for 100 epochs, which takes approximately 32 h on 4 TITAN XP GPUs.

Results: In Fig. 12, we show a few random samples of the generated images. The first column is the sketch map. The second column is the generated image. The third column is the

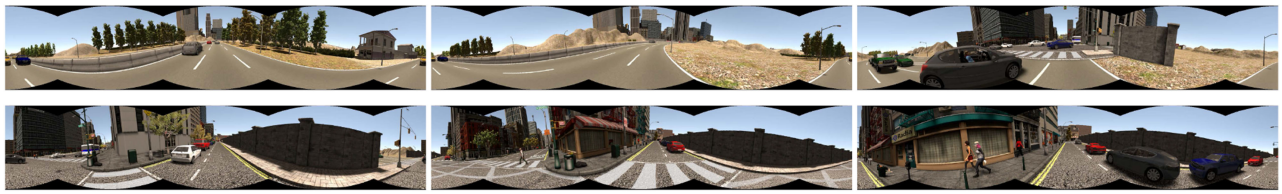


Fig. 11. Some spherical images from SYNTHIA dataset.

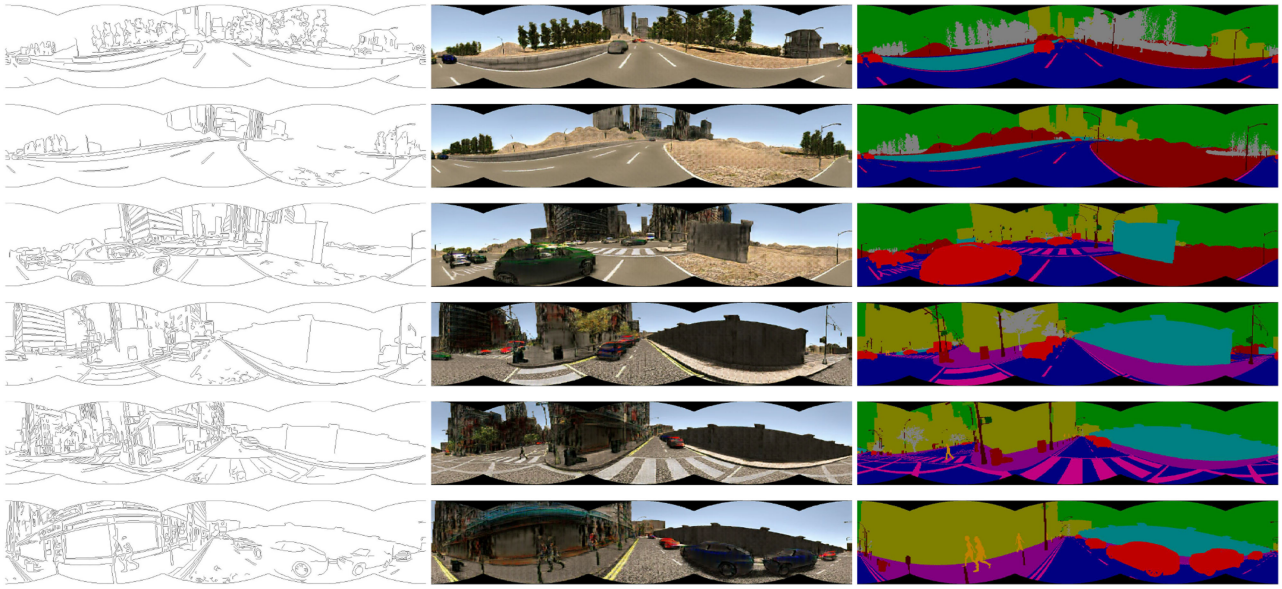


Fig. 12. Visual examples of the images generated on SYNTHIA dataset. From left to right: input sketch map, synthesized image and input label map.

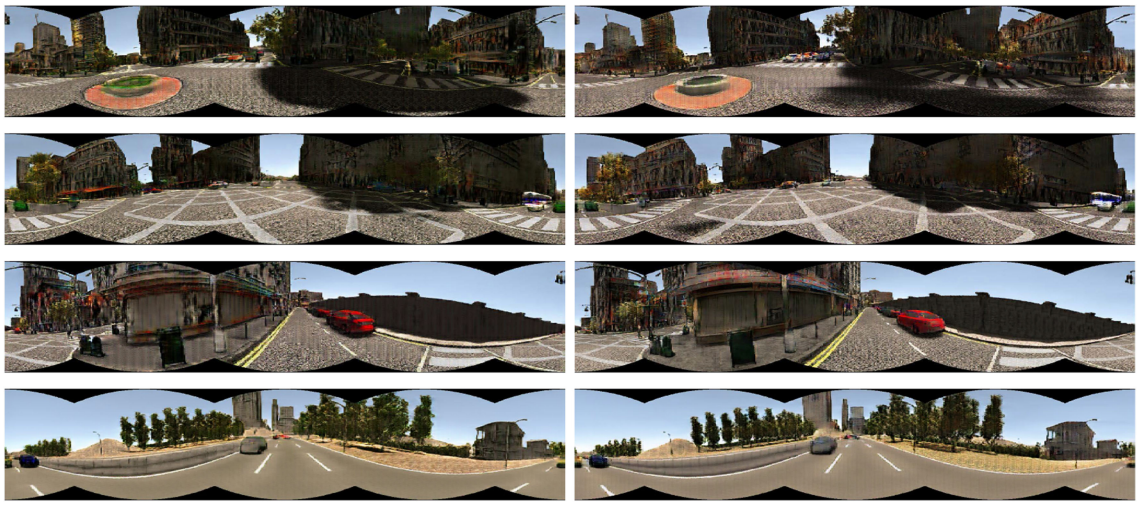


Fig. 13. Visual examples of the generated images with our approach and baseline model on SYNTHIA dataset. From left to right: our approach, pix2pixHD baseline model.

input label map. It is observed that the generated images are often realistic-looking and plausible. Especially the last row, the peoples and pillars are small objects. They are recovered from the sparse inputs clearly. The homogeneous regions, such as road and sky, are also synthesized with high quality. With the input sketch map, more details are preserved. In addition, the label map makes the specified label region more consistent. However, the resolution is slightly worse than that of the original image because all of the inputs are resized to 512×512 px. We resize the generated images to the original size for visualization.

Comparison: Similar to the AOI dataset, we also compare the proposed method with the baseline, the Pix2pixHD

TABLE II
QUANTITATIVE RESULTS ON AOI DATASET. WE REPORT IS SCORES (HIGHER IS BETTER) AND FID SCORES (SMALLER IS BETTER) ON THE TESTING IMAGES

AOI dataset			
method	IS mean	IS std	FID
Ours	1.718	0.208	208.743
Pix2pixHD	1.727	0.380	232.375

TABLE III
QUANTITATIVE RESULTS ON SYNTHIA DATASET. WE REPORT IS SCORES (HIGHER IS BETTER) AND FID SCORES (SMALLER IS BETTER) ON THE TESTING IMAGES

SYNTHIA dataset			
method	IS mean	IS std	FID
Ours	1.965	0.330	80.735
Pix2pixHD	1.895	0.367	84.845

model. The corresponding experimental results are shown in Fig. 13. The first column shows the generated images of our approach. The second column depicts the generated images of the Pix2pixHD model. These two results look comparable. However, the edge portion of the proposed approach is clearer than that of the baseline model, especially the building edges in the third row. It shows that the sketch map improves the quality of the edge portions. Moreover, for the first and the second row, the shadows in the generated image of the baseline model look strange and discontinuous. The phenomenon is avoided in our method because the use of spherical convolution is more reasonable for generating a spherical structure. The above experimental results validate the effectiveness of the proposed approach.

C. Quantitative Evaluations

Objective evaluation: To further evaluate the performance of the proposed approach, we adopt the inception score (IS) [64], [65] and Frechet inception distance (FID) [65], [66] to quantitatively evaluate the performance of the generative models. The IS score reflects the plausibility and variety of the generated images, and a higher IS score represents better results. The FID score captures the distance between the generated images and the real ones. A smaller FID score represents that the generated image is closer to the real image. The numerical indexes of the AOI dataset and SYNTHIA dataset are shown in Table II and Table III, respectively. For the AOI dataset, we compute the IS score and FID score on 56 testing images. The corresponding results are shown in Table II. The IS mean represents the average score on all test images. The IS std represents the standard deviation on all test images. From Table II, we can see that the IS mean of the Pix2pixHD model is slightly higher than that of the proposed approach. We surmise that the phenomenon is caused by the checkerboard artifacts in the Pix2pixHD model. In addition, we need to explain that the IS model is trained on the ImageNet dataset. Its value does not reflect the capabilities of our model well. But we still give the IS scores for reference. FID scores of the proposed approach are 24 points higher than that of the baseline model. For the SYNTHIA dataset, we compute the IS score and FID score on 200 testing images. The numerical indexes are shown in Table III. We can see that both IS scores and FID scores of the proposed approach are superior to those of the baseline model.

TABLE IV
HUMAN EVALUATION RESULTS. WE REPORT THE PREFERENCE % OF OUR RESULTS AND THE PIX2PIXHD RESULTS

Preference of our results	Preference of pix2pixHD results
66%	34%

Subjective evaluation: Moreover, we show the quantitative evaluations of the generated images from a human perspective. Specifically, we invited 30 human evaluators to compare the generated images from the proposed approach and the Pix2pixHD method. For each sample, there are two generated images and the human evaluator selects the better one from the two. We prepared 30 generated image pairs from the generated images of the AOI dataset and SYNTHIA dataset. Each pair includes the generated images from the proposed approach and the baseline model. The corresponding results are shown in Table IV. We can see that the baseline model only wins 34%, while our model wins 66%. Our model is better than the baseline model over by 30% from the human views. The results suggest that the generated images implementing our proposed approach have higher quality.

D. Ablation Experiments

To analyze the effects of different components, we report the IS scores and FID scores on the AOI dataset with a few ablations over the proposed model. To show the experimental results more clearly, the full model is defined as $GAN + SC + L_{FM} + L_{VGG}$.

Spherical convolution: The spherical convolution is mainly used to capture the rotational invariance and multiple angles during the training process. It can reduce the distortion caused by 2-D planar mapping. In fact, the difference between the proposed approach and the baseline model is the spherical convolution. The effect of the spherical convolution is shown in Table II and Table III.

Feature matching loss: The feature matching loss is used to match the intermediate representation from the real and the generated images. To observe the effect, we retrain a model without L_{FM} and define it as $GAN + SC + L_{VGG}$.

VGG loss: VGG loss is a kind of perceptual loss and is used to compute the similarity in the representation space. We also retrain a model without L_{VGG} and define it as $GAN + SC + L_{FM}$.

The corresponding experimental results are shown in Table V. From Table V, we can see that without spherical convolutional the IS scores are almost the same as those in the proposed approach. However, the FID scores of the proposed approach are higher than that of the $GAN + L_{FM} + L_{VGG}$ model. It indicates that the spherical convolution can reduce the distortion in the panoramic image generation. When removing the feature matching loss, IS scores rise. This rise may be caused by the checkerboard artifacts. FID scores have risen markedly. This outcome shows that feature matching loss has a great impact on

TABLE V
ABLATION EXPERIMENTS ON AOI DATASET. WE REPORT IS SCORES (HIGHER IS BETTER) AND FID SCORES (SMALLER IS BETTER) OF THE GENERATED IMAGES

AOI dataset			
methods	IS mean	IS std	FID
$GAN + SC + L_{FM} + L_{VGG}$ (Ours)	1.718	0.208	208.743
$GAN + L_{FM} + L_{VGG}$ (Pix2pixHD)	1.727	0.380	232.375
$GAN + SC + L_{VGG}$	1.867	0.365	265.549
$GAN + SC + L_{FM}$	1.686	0.322	254.708

the image generation. We also provide the performance without VGG loss. We can see that the IS scores drop and the FID scores rise compared to the proposed approach. It demonstrates that adding perceptual loss enhances the results.

V. CONCLUSION

This paper has presented a panoramic image generation approach. Under the GANs framework, we design the generator with spherical convolution to preserve the fidelity of the panoramic image. Meanwhile, a high-resolution panoramic image is generated by only using the sparse sketch map as input. It is a significant improvement from the point of view of communication. This improvement means that a large amount of data is transmitted with very little bandwidth. In addition, we combine GAN loss, feature matching loss and VGG loss as the objective function. The experimental results show that the proposed approach is able to generate the high-resolution panoramic images that have a looking-realistic and plausible quality. Our approach can be used in several applications. For example, in communication systems, the transmitted data comprise the sketch map. It will greatly save the communication bandwidth.

In the future, we will address the following two problems. The panoramic images are generated based on high-level semantics and structures, thereby achieving high visual quality. However, pixel-wise quality is important in some applications. Therefore, it is necessary to combine the high-level semantics and the low-level features to balance the subjective quality and objective quality. Moreover, the spherical convolution needs inputs with a square shape and images of low resolution. Meanwhile, it needs a huge amount of memory. These challenges limit the ability of the model. Additionally, optimizing the spherical convolution and making the model more general are also our future work.

REFERENCES

[1] M. Xu, Y. Song, J. Wang, M. Qiao, L. Huo, and Z. Wang, "Predicting head movement in panoramic video: A deep reinforcement learning approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2593–2708, Nov. 2019.

[2] J. Lu, Y. Yang, R. Liu, S. B. Kang, and J. Yu, "2d-to-stereo panorama conversion using gan and concentric mosaics," *IEEE Access*, vol. 7, pp. 23187–23196, 2019.

[3] M. Xu, C. Li, Z. Chen, Z. Wang, and Z. Guan, "Assessing visual quality of omnidirectional videos," *IEEE Trans. Circuits Syst. for Video Technol.*, vol. 29, no. 12, pp. 3516–3530, Dec. 2019.

[4] G. K. Wallace, "The JPEG still picture compression standard," *IEEE Trans. Consum. Electron.*, vol. 38, no. 1, pp. 18–34, Feb. 1992.

[5] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[6] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 586–597, Dec. 2007.

[7] I. J. Goodfellow *et al.*, "Generative adversarial networks," *Adv. Neural Inf. Process. Syst.*, vol. 3, pp. 2672–2680, 2014.

[8] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vision*, Oct. 2017, pp. 2242–2251.

[9] T. Dekel, C. Gan, D. Krishnan, C. Liu, and W. T. Freeman, "Sparse, smart contours to represent and edit images," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, 2018, pp. 3511–3520.

[10] T. Kawanishi, K. Yamazawa, H. Iwasa, H. Takemura, and N. Yokoya, "Generation of high-resolution stereo panoramic images by omnidirectional imaging sensor using hexagonal pyramidal mirrors," in *Proc. 14th Int. Conf. Pattern Recognit.*, Aug. 1998, vol. 1, pp. 485–489.

[11] R. Mukundan, S. H. Ong, and P. A. Lee, "Image analysis by Tchebichef moments," *IEEE Trans. Image Process.*, vol. 10, no. 9, pp. 1357–1364, Sep. 2001.

[12] S. M. Elshoura and D. B. Megherbi, "Analysis of noise sensitivity and reconstruction accuracy of Tchebichef moments," in *Proc. IEEE Southeastcon*, 2008, pp. 521–526.

[13] Y. L. Hong, Il K. Eom, and Y. S. Kim, "Adaptive reconstruction of lost block using difference of DC in DCT domain," in *Proc. Int. Conf. Signal Process.*, 2004, pp. 853–856.

[14] B. K. Gunturk, Y. Altunbasak, and R. M. Mersereau, "Super-resolution reconstruction of compressed video using transform-domain statistics," *IEEE Trans. Image Process.*, vol. 13, no. 1, pp. 33–43, Jan. 2004.

[15] M. Antonini, P. Mathieu, M. Barlaud and I. Daubechies, "Image coding using wavelet transform," *IEEE Trans. Image Process.*, vol. 1, no. 2, pp. 205–220, Apr. 1992.

[16] J. L. Starck, A. Bijaoui, B. Lopez, and C. Perrier, "Image reconstruction by the wavelet transform applied to aperture synthesis," *Astron. Astrophys.*, vol. 283, no. 283, pp. 349–360, 1994.

[17] C. Bhattacharya and P. R. Mahapatra, "A discrete wavelet transform approach to multiresolution complex sar image generation," *IEEE Geosci. Remote Sens. Lett.*, vol. 4, no. 3, pp. 416–420, Jul. 2007.

[18] Y. Ming and L. Yi, "Model selection and estimation in regression with grouped variables," *J. Roy. Statist. Soc.*, vol. 68, no. 1, pp. 49–67, 2006.

[19] J. A Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.

[20] J. A Tropp and A. C Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.

[21] A. Cohen, W. Dahmen, and R. Devore, "Compressed sensing and best ??-term approximation," *J. Amer. Math. Soc.*, vol. 22, no. 1, pp. 211–231, 2009.

[22] S. Osher, Y. Mao, B. Dong, and W. Yin, "Fast linearized Bregman iteration for compressive sensing and sparse denoising," *Math. Comput.*, vol. 8, no. 1, pp. 93–111, 2010.

[23] Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang, "A survey of sparse representation: Algorithms and applications," *IEEE Access*, vol. 3, pp. 490–530, 2017.

[24] S. G. Mallat and Z. Zhifeng, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[25] F. Hong and H. Yang, "A new compressed sensing-based matching pursuit algorithm for image reconstruction," in *Proc. Int. Congr. Image Signal Process.*, 2013, pp. 338–342.

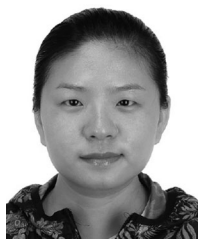
[26] S. Ugur, O. Arikan, and A. C. Grbz, "Sar image reconstruction by expectation maximization based matching pursuit.," *Digit. Signal Process.*, vol. 37, no. 1, pp. 75–84, 2015.

[27] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.

[28] H. S. Goklani, J. N. Sarvaiya, and A. M. Fahad, "Image reconstruction using orthogonal matching pursuit algorithm," in *Proc. Int. Conf. Emerg. Technol. Trends Electron.*, 2015, pp. 1–5.

[29] Meenakshi and S. Budhiraja, "Image reconstruction using modified orthogonal matching pursuit and compressive sensing," in *Proc. Int. Conf. Comput.*, 2015, pp. 1073–1078.

[30] Z. Lin, "Image recovery based on compressive sensing and curvelet transform via romp," in *Proc. 9th Int. Conf. Fuzzy Syst. Knowl. Discovery*, 2012, pp. 1812–1815.
[31] T. T. Do, L. Gan, N. Nguyen, and T. D. Tran, "Sparsity adaptive matching pursuit algorithm for practical compressed sensing," in *Proc. Conf. Signals, Syst. Comput.*, 2010, pp. 581–587.
[32] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Appl. Comput. Harmon. Anal.*, vol. 27, no. 3, pp. 265–274, 2009.
[33] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, "Bregman iterative algorithms for l-minimization with applications to compressed sensing," *SIAM J. Imag. Sci.*, vol. 1, no. 1, pp. 143–168, 2008.
[34] K. Egiazarian, A. Foi, and V. Katkovnik, "Compressed sensing image reconstruction via recursive spatially adaptive filtering," in *Proc. IEEE Int. Conf. Image Process.*, 2007, vol. 1, pp. 549–552.
[35] W. Dong, L. Zhang, G. Shi, and X. Wu, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization.," *IEEE Trans. Image Process.*, vol. 20, no. 7, pp. 1838–1857, Jul. 2011.
[36] G. Yu, G Sapiro, and S Mallat, "Solving inverse problems with piecewise linear estimators: From Gaussian mixture models to structured sparsity," *IEEE Trans. Image Process.*, vol. 21, no. 5, pp. 2481–2499, May 2012.
[37] W. Dong, Z. Lei, and G. Shi, "Centralized sparse representation for image restoration," in *Proc. IEEE Int. Conf. Comput. Vision*, 2011, pp. 1259–1266.
[38] W. Dong, L. Zhang, G. Shi, and X. Li, "Nonlocally centralized sparse representation for image restoration," *IEEE Trans. Image Process.*, vol. 22, no. 4, pp. 1620–1630, Apr. 2013.
[39] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2014.
[40] T. J. OShea, T. Roy, and T. C. Clancy, "Over-the-air deep learning based radio signal classification," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 1, pp. 168–179, Feb. 2018.
[41] A. van den Oord, N. Kalchbrenner, L. Espeholt, K. kavukcuoglu, O. Vinyals, and A. Graves, "Conditional image generation with pixelcnn decoders," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds., New York, NY, USA: Curran Associates, Inc., 2016, pp. 4790–4798.
[42] S. Lohit, K. Kulkarni, R. Kerviche, P. Turaga, and A. Ashok, "Convolutional neural networks for non-iterative reconstruction of compressively sensed images," *IEEE Trans. Comput. Imag.*, vol. 4, no. 3, pp. 326–340, Sep. 2018.
[43] Y. Li, W. Xie, and H. Li, "Hyperspectral image reconstruction by deep convolutional neural network for classification," *Pattern Recognit.*, vol. 63, pp. 371–383, 2017.
[44] Y. Rivenson, Y. Zhang, H. Günaydin, Da Teng, and A. Ozcan, "Phase recovery and holographic image reconstruction using deep learning in neural networks," *Light Sci. Appl.*, vol. 7, no. 2, pp. 17141–17141, 2018.
[45] B. Kelly, T. P. Matthews, and M. A. Anastasio, "Deep learning-guided image reconstruction from incomplete data," 2017, *arXiv:1709.00584*.
[46] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning what and where to draw," in *Proc. Advances Neural Inf. Process. Syst.*, 2016, pp. 217–225.
[47] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 214–223.
[48] J. Zhao, M. Mathieu, and Y. Lecun, "Energy-based generative adversarial network," 2016, *arXiv:1609.03126*.
[49] P. Isola, J. Yan Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 5967–5976.
[50] T. Wang, M. Liu, J. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, Jun. 2018, pp. 8798–8807.
[51] H. B. Barlow, "Vision: A computational investigation into the human representation and processing of visual information: David Marr. San Francisco: W. H. Freeman, 1982. pp. xvi + 397," *J. Math. Psychol.*, vol. 27, no. 1, pp. 107–110, 1983.
[52] C. En Guo, S. Chun Zhu, and N. Wu Ying, "Primal sketch: Integrating structure and texture," *Comput. Vis. Image Understand.*, vol. 106, no. 1, pp. 5–19, 2007.
[53] J. Wu, F. Liu, L. Jiao, X. Zhang, H. Hao, and S. Wang, "Local maximal homogeneous region search for Sar speckle reduction with sketch-based geometrical Kernel function," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 9, pp. 5751–5764, Sep. 2014.
[54] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jul. 2017, pp. 5967–5976.
[55] T. Cohen and M. Welling, "Group equivariant convolutional networks," in *Proc. Int. conf. mach. learn.*, 2016, pp. 2990–2999.
[56] T. S. Cohen, M. Geiger, J. Koehler, and M. Welling, "Spherical CNNs," https://arxiv.org/abs/1801.10130, 2018.
[57] S. Santurkar, D. Budden, and N. Shavit, "Generative compression," in *Proc. IEEE Picture Coding Symp.*, 2018, pp. 258–262.
[58] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds., New York, NY, USA: Curran Associates, Inc., 2016, pp. 658–666.
[59] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jul. 2017, pp. 105–114.
[60] M. Xu, L. Yang, X. Tao, Y. Duan, and Z. Wang, "Saliency prediction on omnidirectional images with generative adversarial imitation learning," 2019, *arXiv:1904.07080*.
[61] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2016, pp. 3234–3243.
[62] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.
[63] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. V. Gool, "Generative adversarial networks for extreme learned image compression," in *Proc. IEEE Int. Conf. Comput. Vision*, 2019, pp. 221–231.
[64] T. Salimans *et al.*, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds., New York, NY, USA: Curran Associates, Inc., 2016, pp. 2234–2242.
[65] W. Xu, S. Keshmiri, and G. R. Wang, "Adversarially approximated autoencoder for image generation and manipulation," *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2387–2396, Sep. 2019.
[66] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., New York, NY, USA: Curran Associates, Inc., 2017, pp. 6626–6637.



Yiping Duan (Member, IEEE) received the B.S. degree from the School of Computer Science and Technology, Henan Normal University, Xinxiang, China, in 2010, and the Ph.D. degree from the School of Computer Science and Technology, Xidian University, Xi'an, China, in 2016. She is currently an Assistant Research Fellow with the Department of Electronic Engineering, Tsinghua University, Beijing, China. Her current research interests include wireless communications, machine learning, computer vision and image processing.

Chaoyi Han received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China in 2016, where he is currently working toward the Ph.D. degree. His current research interests include image processing and machine learning.

Xiaoming Tao (Member, IEEE) received the B.E. degree from the school of Telecommunications Engineering, Xidian University, Xi'an, China, in 2003, and the Ph.D. degree from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2008. She is currently a Professor with the Department of Electronic Engineering, Tsinghua University. Her research interests include wireless communications and networking, and multimedia signal processing.

Bingrui Geng received the B.E. degree from the School of Mechano-Electronic Engineering, Xidian University, Xi'an, China, in 2011, and the Ph.D. degree from the School of Electronic Engineering, Xidian University, Xi'an, China, in 2018. She is currently working toward Postdoctoral Research with the Department of Electronic Engineering, Tsinghua University, Beijing, China. Her current research interests include optimization and network data mining.

Yunfei Du received the B.S. degree in electrical and information engineering from Shaanxi University of Science and Technology, Xi'an, China, in 2017. He is currently working toward the master's degree with Xidian University, Xi'an, China. His current research interests include deep learning and image processing.

Jianhua Lu (Fellow, IEEE) received the B.S.E.E. and M.S.E.E. degrees from Tsinghua University, Beijing, China, in 1986 and 1989, respectively, and the Ph.D. degree in electrical and electronic engineering from the Hong Kong University of Science and Technology, Hong Kong. Since 1989, he has been with the Department of Electronic Engineering, Tsinghua University, where he serves as a Professor. He has authored more than 180 technical papers in international journals and conference proceedings. His research interests include broadband wireless communication, multimedia signal processing, and wireless networking. He is also a Fellow of the IEEE Communication Society and the IEEE Signal Processing Society. He has served in numerous IEEE conferences as a member of Technical Program Committees and served as the Lead Chair of the General Symposium of IEEE ICC 2008, as well as a Program Committee Co-Chair of the 9th IEEE International Conference on Cognitive Informatics in 2010. He has been an active member of professional societies. He was the recipient of best paper awards at the IEEE International Conference on Communications, Circuits and Systems 2002, ChinaCom 2006, and IEEE Embedded-Com 2012, and the National Distinguished Young Scholar Fund by the NSF Committee of China in 2005. He is now a Chief Scientist of the National Basic Research Program (973), China.