Journal of Network Intelligence ©2024 ISSN 2414-8105
Taiwan Ubiquitous Information Volume **9**, Number **4**, November **2024**

PDEGAN: A Panoramic Style Transfer Based on Generative Adversarial Networks

Qinghua Wang†

College of Computer and Big Data,Minjiang University, Fuzhou
350108, China
Fuzhou Technology Innovation Center of intelligent Manufacturing information System,
Minjiang University, Fuzhou 350108, China
Engineering Research Center for ICH Digitalization and Multi-source Information Fusion
(Fujian Polytechnic Normal University), Fujian Province University, Fuzhou 350300,
China
2595666958@qq.com

Xinling Long†

College of Computer and Control Engineering, Minjiang University, Fuzhou 350108, China
1351707884@qq.com

Jingwei Huang

College of Computer and Big Data, Fuzhou University, Fuzhou, China
huangjw0824@163.com

Yang Chen

School of Mechanical and Automotive Engineering, Fujian University of Technology, Fuzhou, China
chenyang000515@163.com

Lirong Yang

College of foreign languages, Dalian Jiaotong University, Dalian 116028, China
1500787371@qq.com

Fuquan Zhang*

College of Computer and Big Data, Minjiang University, Fuzhou 350108, China
Digital Media Art, Key Laboratory of Sichuan Province, Sichuan Conservatory of Music,
Chengdu 610021, China
Fuzhou Technology Innovation Center of intelligent Manufacturing information System,
Minjiang University, Fuzhou 350108, China
Engineering Research Center for ICH Digitalization and Multi-source Information Fusion
(Fujian Polytechnic Normal University), Fujian Province University, Fuzhou 350300,
China
8528750@qq.com

† These authors contributed equally to this work.
* Corresponding author: Fuquan Zhang

Received April 23, 2023, revised September 26, 2023, accepted December 24, 2023.

Abstract. *Panoramic image plays an extremely important role in the application of 3D, but in some special scenarios, the depth information is also needed as an auxiliary, but to obtain the panoramic depth information needs a high cost, it is extremely difficult to obtain. Therefore, this paper proposes a panoramic depth estimation method called PDEGAN (Panoramic Depth Estimation Generative Adversarial Network). This method first enhances the style migration, transfers the input panoramic image to a style closer to the panoramic depth image domain, and then combines the conversion results with the original panoramic image as the input data for further depth estimation training. By using the style closer to the depth estimation, the texture, color and other characteristics of the image can be adjusted, so as to better highlight the depth information of the object in the image, so that the generator network is more focused on performing the image translation task, reduce the impact of style differences on the image translation task, and improve the generalization ability of the model. At the same time, the model will also use multi-scale feature fusion, combined with the attention mechanism, to better capture the depth information of different scales, viewing angles and channels. Not only that, but also by introducing depth gradient loss, surface normal loss to improve the accuracy and stability of depth estimation. To this end, we also made a set of related datasets for the related experiments of PDEGAN, and named it PDE Images. Finally, it is shown that the method proposed in this paper has more accurate estimation results than the traditional Pix2Pix image conversion model.*
Keywords: panoramic depth estimation, style transfer enhancement, attention mechanism

1. Introduction.

Panoramic technology is a kind of image technology based on the camera or virtual reality technology, the actual environment or virtual scene information can be saved on the image, and presented in a panoramic form. With the rapid development of XR (Extended Reality) technology, panoramic technology has been widely used in many fields such as digital education, marketing, real estate and other [1, 2, 3]. However, in some special scenes, panoramic depth information is also needed as an auxiliary, and panoramic depth image is a common way to store depth information. However, compared to obtaining panoramic images, obtaining its depth information requires extremely professional equipment, which is quite difficult for most people who want to apply panoramic depth information. Therefore, there are common methods divided into two kinds. The first is depth calculation based on geometry [4]. This method mainly calculates the spatial relationship between images from different perspectives to estimate the depth, but this method lacks the use of semantic information of RGB images. The other is the depth estimation based on deep learning method [5], which can take the semantic information of RGB images as the depth estimation factor. Compared with the depth estimation of pure geometry, it can also support single-perspective estimation and reduce the need for input information. Pix2Pix [6] is a common framework based on deep learning in the depth estimation scheme. Entering a single panoramic image, it can be converted into the corresponding panoramic depth image. However, since Pix2Pix is based on GAN [7] framework, it is prone to common model collapse and model instability in the training process. The trained models can only be applied to specific scenes and data distribution, and it is difficult to generalize to other different scenes, resulting in unstable and accurate generated deep image quality. Because the panoramic depth image itself is extremely large, it is more difficult to collect training data for various scenarios. In view of this problem, this paper will convert the input image to the image domain more suitable for depth estimation with the help of style migration enhancement, and add the depth gradient loss and surface normal loss for further constraints, so as to improve the generalization ability and stability of the model. Due to the projection relationship, the image of panoramic type has some regular distortion problem in the storage of information, and the information scales of different positions are different. In the traditional CNN (Convolutional Neural Network) [8], limited by its limited receptive field and fixed sampling location, it is difficult to handle this distortion property, which leads to a decrease in the precision of depth estimation. To address this issue, this paper also incorporates attention mechanisms to improve the accuracy of depth estimation.

2. Method.

2.1. **PDEGAN model.** Figure 1 shows the PDEGAN model's structure as it is presented in this article. The model first provides $G_c : X \to T$ and $F_c : T \to X$ as two corresponding mappings. The input panoramic image domain is X, the panoramic depth image domain is T, and the corresponding discriminators are D_{cT} and D_{cX}, respectively. Following the splicing of the panoramic picture of the input x and the image t produced by the generator G_c as the input of the generator G, the corresponding panoramic depth image y is then produced.

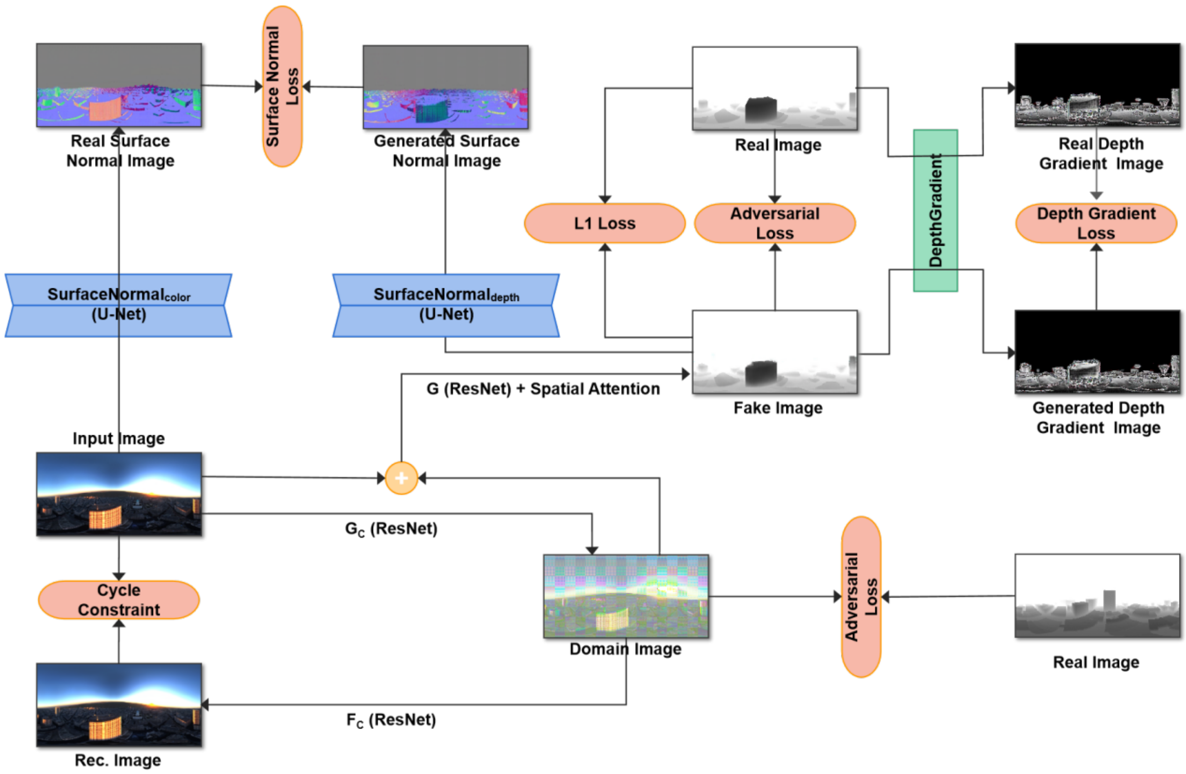

FIGURE 1. PDEGAN

In the style migration enhancement stage, in order to better capture the characteristics of different domains and thus achieve more accurate style migration, ResNet [9] is used as the network model of generators G_c and F_c. In the stage of translating into corresponding panoramic depth images, U-Net [10] is used as the network model of generator G, the main reason is that it can preserve fine local structures and details when generating images, thus achieving more accurate and natural translation results.At the same time, different from the generation of the depth image in the conventional plane, the panoramic depth image is affected by the information projection, and the image on the two-dimensional screen presents some special regular distortion, which makes the panoramic depth estimation task difficult. In this regard, this paper introduces the attention mechanism to solve this problem.

For the panoramic depth image task, more attention is paid to its structure and texture features. Therefore, this paper uses PatchGAN as a network model for all discriminants in PDEGAN [11], which is extremely effective in retaining its content and details. Meanwhile, The following calculation demonstrates how PDEGAN directly constrains G through two common losses, $L_{GAN}(G, G_c, D)$ and $L_{l_1}(G, G_c)$.

$$\mathcal{L}_{GAN}(G, G_c, D) = \mathbb{E}_{x \sim p_{data}(x), y \sim p_{data}(y)}[log(D(x, y))] + \mathbb{E}_{x \sim p_{data}(x)}[log(1 - D(x, G(x, G_c(x))))] \tag{1}$$

$$\mathcal{L}_{l_1}(G, G_c) = \mathbb{E}_{x \sim p_{data}(x), y \sim p_{data}(y)}[||G(x, G_c(c)) - y||] \tag{2}$$

In addition, there are depth gradient constraints, along with surface normal constraints, to further enhance the accuracy of the generated image, which is detailed in the later sections.

2.2. **Enhanced style migration.** Style transfer enhancement refers to transferring the style of one image to another image domain that is more suitable for processing. In the panoramic depth estimation, the results are greatly influenced by the data set. Therefore, PDEGAN converts the input image into an image closer to the target style through style migration, enabling the generator to focus more on performing the image translation task, thus reducing the impact of style differences on the image translation task and improving the generalization ability of the model. This module adds a cycle consistency loss of $L_{cycle}(G_c, F_c)$ to make corresponding constraints, where λ_c is its correlation coefficient. This constraint makes $F_c(G_c(x)) \approx X$ and $G_c(F_c(t)) \approx T$. Therefore, there are two generative confrontation losses, namely $L_{GAN}(G_c, D_{cT}, X, T)$ and $L_{GAN}(G_c, D_{cT}, X, T)$, and the final loss $L_{transition}(G_c, F_c, D_{cT}, D_{cX})$ of this part is as follows:

$$\mathcal{L}_{cycle}(G_c, F_c) = \mathbb{E}_{x \sim p_{data}}(x)[||F_c(G_c(x)) - x||_1] + \mathbb{E}_{t \sim p_{data}}(t)[||G_c(F_c(t)) - t||_1] \tag{3}$$

$$\mathcal{L}_{GAN}(G_c, D_{cT}, X, T) = \mathbb{E}_{t \sim p_{data}(t)}[log(D_{cT}(t))] + \mathbb{E}_{x \sim p_{data}(x)}[log(1 - D_{cT}(G_c(x)))] \tag{4}$$

$$\mathcal{L}_{GAN}(F_c, D_{cX}, T, X) = \mathbb{E}_{x \sim p_{data}(x)}[log(D_{cX}(x))] + \mathbb{E}_{t \sim p_{data}(t)}[log(1 - D_{cX}(F_c(t)))] \tag{5}$$

$$\mathcal{L}_{transition}(G_c, F_c, D_{cT}, D_{cX}) = \lambda_c \mathcal{L}_{cycle}(G_c, F_c) + \mathcal{L}_{GAN}(G_c, D_{cT}, X, T) + \mathcal{L}_{GAN}(F_c, D_{cX}, T, X) \tag{6}$$

2.3. **Attention mechanism.** In the panoramic depth image, there are obvious differences in the information scale of different regions. This will result in the generative model sensing different regions, thus affecting the accuracy of depth estimation. Therefore, this paper introduces CBAM (Convolutional Block Attention Module) [12] as its attention module, which contains two key parts, which are channel attention and spatial attention. The module is located at the top of the generator G and uses channel attention to weight the stitching results of the panoramic image in the previous step of the model. This adaptive computational approach contributes to learning the importance of different channels in the task. Its spatial attention enables the model to focus on different regions in a more reasonable way by weighting calculation on the spatial dimension, so as to improve the perception of different regions. In addition to including attention, CBAM also includes feature fusion, which integrates the information in different feature maps to get a more comprehensive feature expression, thus improving the accuracy of depth estimation.

2.4. **Depth gradient constraint.** In order to strengthen the network of the local changes of the object edge and surface, so as to improve the robustness and accuracy of depth estimation, this paper through a pre-trained U-Net as a depth gradient extractor: DepthGradlent performs depth-gradient-constrained tasks, the depth gradient constraint focus on the edge of the depth image information and local changes, it helps to capture the object edge and surface details. The reason for using U-Net here is its better flexibility and better performance for deep gradient computing that wants to use some more complex methods. In panoramic depth estimation, depth gradient constraints can improve prediction accuracy in edge regions, especially if there is significant depth variation at the edge of an object. In the depth gradient constraint, the loss $L_{depth\,gradlent}(G, G_c)$ is as follows:

$$\mathcal{L}_{depth\,gradlent}(G, G_c) = \mathbb{E}_{x \sim p_{data}(x)}[||DepthGradlent(G(x, G_c(x))) - DepthGradlent(x)||_1] \tag{7}$$

2.5. **Surface normal constraint.** To better capture the geometry and surface structure of the object. When the surface of an object has a complex texture or obvious concave and convex changes, the surface normal constraints can provide better recovery of surface details. The normal constraint guides network learning by converting the predicted panoramic depth map into a panoramic surface normal graph and calculating the difference between it and the real panoramic surface normal graphs. This constraint helps the network to focus on the normal direction of the object surface during training, thus improving the accuracy of panoramic depth estimation. In this paper, two pre-trained U-Net are used as the surface normal extractors: $SurfaceNormal_{color}$ and $SurfaceNormal_{depth}$, the function of $SurfaceNormal_{color}$ is to extract the panoramic surface normal data from the panoramic image, but also $SurfaceNormal_{depth}$ is to extract the panoramic surface normal data from the panoramic depth image.The reason for using U-Net here is that it can retain fine local structure and details when generating images, so as to achieve more accurate and natural translation results, and better complete the generation task of panoramic surface normal images. In the surface normal constraint, the loss $L_{surface\,normal}(G, G_c)$ is as follows:

$$\mathcal{L}_{surface\,normal}(G, G_c) = \mathbb{E}_{y \sim p_{data}(y)}[||SurfaceNormal(G(x, G_c(x))) - SurfaceNormal(x)||_1] \tag{8}$$

2.6. **Global loss.** By integrating the individual modules, the overall loss $\mathcal{L}(G, F, D_Y, D_X, D_{blur})$ of the final FCIGAN is as follows:

$$\begin{aligned}\mathcal{L}_{(G,F,D_Y,D_X,D_{blur})} = \alpha\mathcal{L}_{transition}(G_c, F_c, D_{cT}, D_{cX}) \quad &+ \beta\mathcal{L}_{depth\,gradient}(G, G_c) \\ + \gamma\mathcal{L}_{surface\,normal}(G, G_c) \quad + \delta\mathcal{L}_{GAN}(G, G_c, D) \quad &+ \varepsilon\mathcal{L}_{l_1}(G, G_c)\end{aligned}$$

Among them, $mathcalL_{transition}(G_c, F_c, D_{cT}, D_{cX})$, $mathcalL_(depthgradlent)(G, G_c)$, $mathcalL_{surface\,normal}(G, G_c)$, $mathcalL_{GAN}(G, G_c, D)$ and $mathcalL_{l_1}(G, G_c)$ are style transfer enhancement loss, panoramic depth gradient loss, panoramic surface normal loss, generation confrontation loss and Manhattan norm loss, $\alpha, \beta, \gamma, \delta$ and ε are style transfer enhancement loss, panoramic depth gradient loss, The weight coefficients of panoramic

TABLE 1. Computer-related configurations used in the experiment

OS	Ubuntu 20.04.4 LTS (64 bit)
CPU	Intel® Xeon(R) Silver 4210 CPU @ 2.20GHz × 40
GPU	NVIDIA GeForce RTX 3080 Ti (12GB) × 4
RAM	128 GB

surface normal loss, generative adversarial loss, and Manhattan norm loss in the overall loss.

3. Results and Discussions.

3.1. Experimental configuration. Computer-related configurations used in this experimental training are shown in Table 1:

The software environment is the PyTorch (1.13.0 + cu116) deep learning framework and the Python 3.8 programming environment. The learning rate of both generator and discriminator was 0.0002, and Adam optimizer was used for the adaptive learning rate. Training Epoch times is 200 and Batch Size is 1. The pre-trained feature extractor has all trained Epoch times of 200. To better capture the details in the image, all discriminators in this method are PatchGAN builds of 30×62 (Height×Width).

3.2. Dataset description. This project collected four types of 512 px×256 px (Width×Height) size data sets, namely, panorama data set, panoramic depth map data set, panoramic depth gradient map data set and panoramic surface normal graph data set. We named it PDE Images, and its various types are shown in Figure 2, and each type of data contains 1000 training sets and 200 test sets.

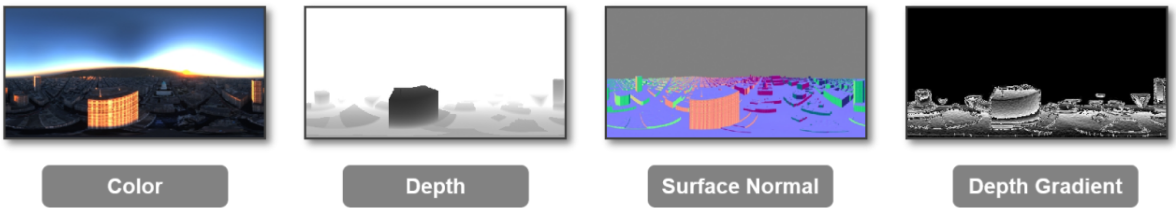

FIGURE 2. PDE Images

The 3D rendering model used in the production of the data set came from the Internet ,and used blender software to carry out the relevant rendering work to generate four different types of image data. By using Blender software, we can render these models with high quality to produce realistic images.

3.3. Evaluation indicators. The SSIM (Structural Similarity Index Measure) [13] and PSNR (Peak Signal To Noise Ratio) [14] were used for comprehensive assessment. SSIM is a hypothesis that extracts information based on the perception of the image. It comprehensively takes into account many factors including the brightness, contrast and structure of the image. The results range between 0 and 1, and the closer the value to 1 indicates the higher the image similarity. The calculation formula is shown as follows:

$$l(x,y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \tag{9}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \tag{10}$$

$$s(x,y) = \frac{\sigma_{xy} + c_2}{\sigma_x\sigma_y + c_2} \tag{11}$$

$$SSIM(x,y) = l(x,y)^{\alpha} c(x,y)^{\beta} s(x,y)^{\gamma} \tag{12}$$

$l(x,y)$ is the brightness based on the mean estimate, $c(x,y)$ is the contrast based on the variance estimate, and $s(x,y)$ is the structural similarity based on the covariance estimate. $c_1 = (K_1L)^2, c_2 = (K_2L)^2, c_3 = c_2/2$, The main role of, is to avoid the instability caused by the formula close to 0, where, K_1 and K_2 are two constants, l is the series of pixel value. Finally, α, β, and γ are set to 1, the calculation process of the final SSIM value is as follows:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{13}$$

PSNR is a method that can be used to measure the ratio between the effective information and the noise of an image. It can effectively reflect whether the image is distorted. The larger the PSNR value, the better the quality of the image. The calculation process is shown as follows:

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2 \tag{14}$$

$$MAX_l = 2^n - 1 \tag{15}$$

$$PSNR = 10\,ln\left(\frac{MAX_I^2}{MSE}\right) \tag{16}$$

Where, MSE is the mean square error between the target image I and the noisy image K, n is the number of digits represented by the image value using binary, and MAX_I is the maximum possible pixel value of the image.

3.4. **Results analysis.** To verify the superiority of the proposed method in this paper, corresponding experiments were conducted with PDEGAN and Pix2Pix, respectively, and Figure 3 shows the images generated by the test set through the two models. It can be found that the estimated results of Pix2Pix and the real panoramic depth image are very different, while PDEGAN performs relatively well. This shows that our proposed method has higher accuracy and fidelity in image generation.

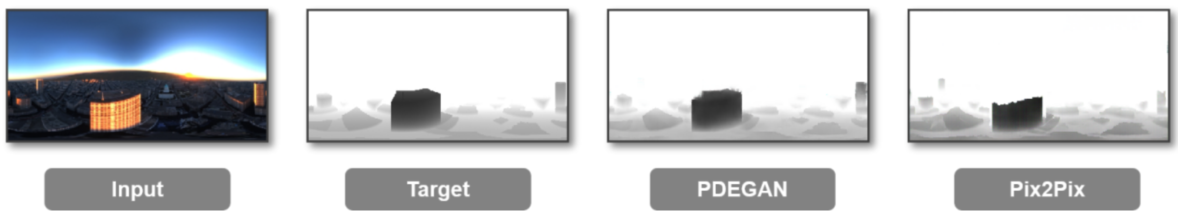

Figure 3. The contrast effects of the test set in PDEGAN versus Pix2Pix

To more effectively illustrate the superiority of PDEGAN over Pix2Pix. we conducted further evaluation. In this experiment, the training set and test set were used as input data for SSIM and PSNR evaluation, and the final corresponding evaluation results are shown in Figure 4 and Figure 5, which can clearly show the performance advantages of PDEGAN over pix2pix in these indicators.

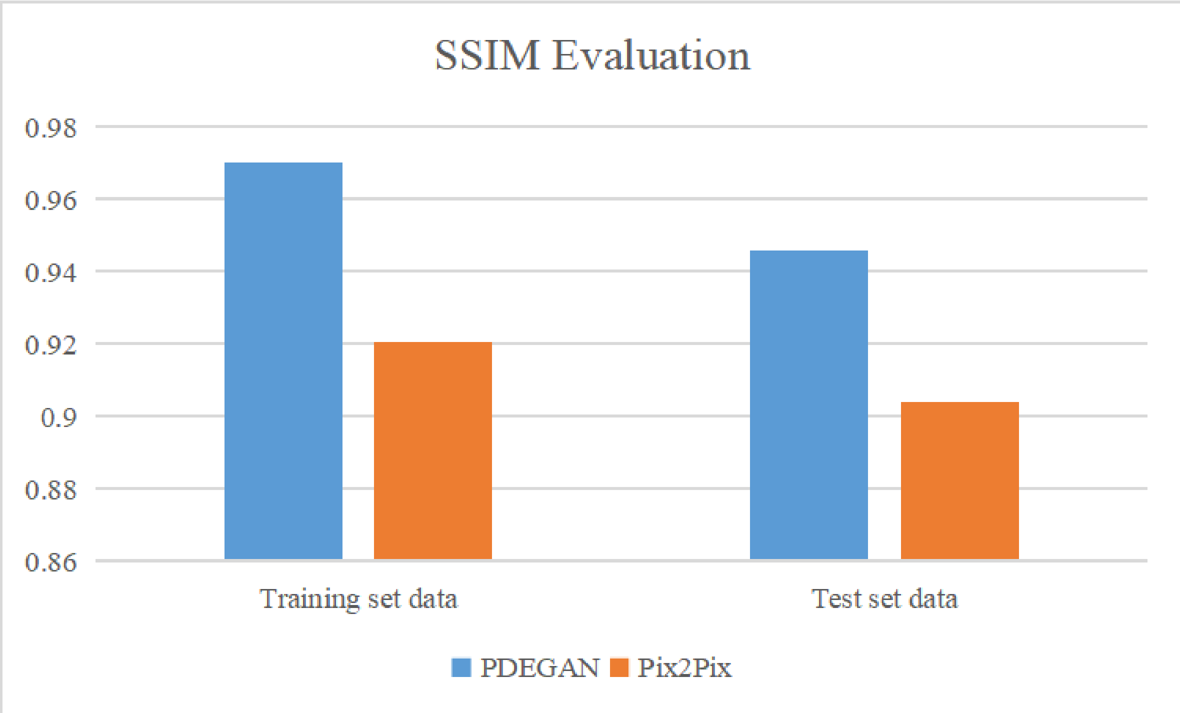

FIGURE 4. SSIM Evaluation

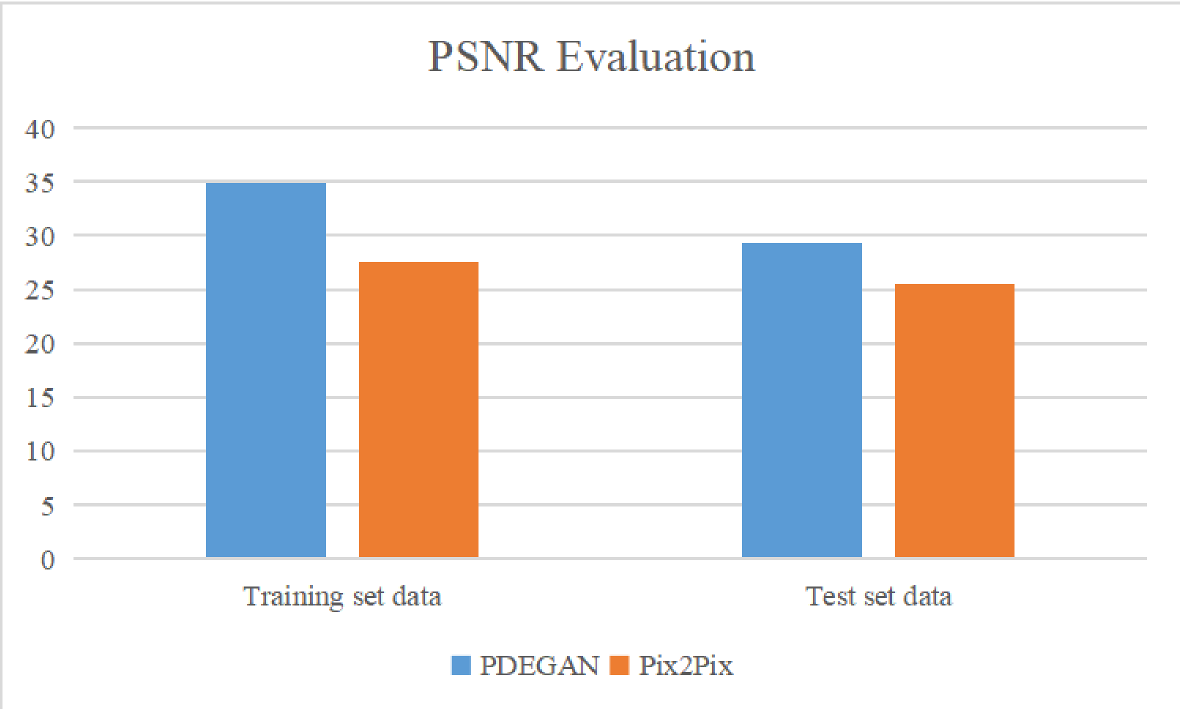

FIGURE 5. PSNR Evaluation

After careful comparative analysis, we can clearly observe that the PDEGAN method proposed in this paper shows a better performance than the Pix2Pix algorithm, both in terms of SSIM evaluation and PSNR evaluation. In SSIM evaluation, the image generated by PDEGAN has higher structural similarity with the real panoramic depth image, indicating that PDEGAN can better maintain the structure and details of the image. In PSNR evaluation, the image generated by PDEGAN has a higher peak signal-to-noise ratio, indicating that PDEGAN can more accurately restore the definition and quality of the image. These evaluation results further verify the superiority of PDEGAN method in image generation task.

4. **Conclusion.** In this article, we explore the challenges of panoramic technology in depth estimation, address the limitations of existing methods, and propose a generative adversarial network model PDEGAN for panoramic depth estimation based on style migration enhancement and attention mechanism. The model combines two phases of style migration and generating panoramic depth images. First, the two mapping relationships are used to convert the input image to the image domain closer to the target, and then the panoramic image and the corresponding style migration enhanced image are stitched together for further panoramic depth image generation as the input data. In the style migration enhancement phase, we adopted ResNet as a generator network model to achieve more accurate style migration. During the generation panoramic depth image phase, we adopted U-Net as the generator network model to retain fine local structure and details. This method not only improves the accuracy and stability of the model in the generation of panoramic depth image, but also improves the generalization ability of the model. At the same time, the model combined with attention mechanism can more effectively capture the depth information of different scales, perspectives and channels through CBAM. In addition, the depth gradient loss and surface normal loss are added to restrict the model more strictly to improve the accuracy of the model. In order to prove the effectiveness of the proposed method through experiments, we also collected and created the corresponding dataset PDE Images, and finally proved that PDEGAN performs well compared with the traditional depth estimation network through SSIM and PSNR. Therefore, the research in this paper is important in driving the field of panoramic depth estimation.

Funding statement. This research was financially supported by the National Innovation and Entrepreneurship Training Project for University (China) Program (2022) under Grant No. 202210395009, and the Minjiang University President's Fund Project (2022) under Grant No. 103952022116. Supported by the project of Digital Media Art, Key Laboratory of Sichuan Province (Sichuan Conservatory of Music, Project No. 21DMAKL01), the first batch of industry-university cooperation collaborative education project funded by the Ministry of Education of the People's Republic of China (Minjiang University, Project No. 202101071001), Minjiang University 2021 school-level scientific research project (Minjiang University, Project No. MYK21011), Open Fund Project of Fuzhou Technology Innovation Center of Intelligent Manufacturing Information System (Minjiang University, Grant No. MJUKF-FTICIMIS2022), Open Fund Project of Engineering Research Center for ICH Digitalization and Multi-source Information Fusion (Fujian Polytechnic Normal University, Grant No. G3-KF2204), Guiding Project of Fujian Province (Minjiang University, Project No. 2020H0046), Key Technology Research and Industrialization Project for Software Industry Innovation in Fujian Province (MinjiangUniversity and Fujian Guotong Information Technology Co., Ltd., Project No. 36). The authors also wish to acknowledge the Department of Science and Technology of Fujian Province

Project: Research on the Model and Algorithm of 3D Clothing Style Transfer for Point Cloud Data Machine Learning (2023J011406).

REFERENCES

[1] C. Snelson, and Y.-C. Hsu, "Educational 360-degree videos in virtual reality: A scoping review of the emerging research," *TechTrends*, vol. 64, no. 3, pp. 404–412, 2020.

[2] K.-A. Ritter, and T.-L. Chambers, "Three-dimensional modeled environments versus 360 degree panoramas for mobile virtual reality training," *Virtual Reality*, vol. 26, no. 2, pp. 571–581, 2022.

[3] R. Eiris, M. Gheisari, and B. Esmaeili, "PARS: Using augmented 360-degree panoramas of reality for construction safety training," *International Journal of Environmental Research and Public Health*, vol. 15, no. 11, pp. 2452, 2018.

[4] R. Zabih, and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Computer Vision—ECCV'94: Third European Conference on Computer Vision Stockholm*. IEEE, 1994, pp. 151–158.

[5] T. Zhou, M. Brown, and N. Snavely, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 1851–1858.

[6] P. Isola, J.-Y. Zhu, and T. Zhou, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 1125–1134.

[7] A. Creswell, T. White, and V. Dumoulin, "Generative adversarial networks: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[8] Y. LeCun, L. Bottou, and Y. Bengio, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[9] K. He, X. Zhang, and S. Ren, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 770–778.

[10] K.-M. He, X.-Y. Zhang, S.-Q. Ren, and J. Sun, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference*. IEEE, 2015, pp. 234–241.

[11] J.-Y. Zhu, T. Park, and P. Isola, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2017, pp. 2223–2232.

[12] S. Woo, J. Park, J.-Y. Lee, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision*. IEEE, 2018, pp. 3–19.

[13] W. Zhou, B. Alan and S. Hamid, and S. Eero, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[14] T. Patrick, and H. David, "Perceptual image distortion," in *Proceedings of 1st International Conference on Image Processing*. IEEE, 1994, pp. 982–986.