

RESEARCH ARTICLE

Where2Stand: Toward a Framework for Portrait Position Recommendation in Photography

ZHENQUAN SHI¹, QINGGANG HOU², GUANJUN SHENG², YONGZHEN KE^{2,3},
KAI WANG², AND YUNGANG JIA⁴

¹School of Software Engineering, Tiangong University, Tianjin 300387, China

²School of Computer Science and Technology, Tiangong University, Tianjin 300387, China

³National Demonstration Center for Experimental Engineering Training Education, Tianjin 300387, China

⁴Tianjin Branch of National Computer Network and Information Security Management Center, Tianjin 300387, China


Corresponding author: Yongzhen Ke (keyongzhen@tiangong.edu.cn)

ABSTRACT Composition layout is crucial in the portrait location recommendation of photography. The existing studies require both landscape background images and portrait foreground images, which limits the scope of practical applications. In this paper, we propose an end-to-end portrait location recommendation model, which mainly consists of three sub-networks: the first sub-networks is the portrait generation network, which generates relatively real portrait foreground images based on random input noise; the second sub-networks is the spatial transformation network, which mainly changes the size and location of the generated portrait based on the input landscape image; The third sub-networks is the compose network to generate a realistic portrait landscape image, which considers not only the correlation between the portrait foreground and the landscape background but also the overall composition aesthetics. Last, the proper standing position is obtained by computing the difference between the generated and input landscape images. We also construct a portrait landscape photo dataset PLDataset to train and verify our method. The experimental results on our dataset show that our proposed method can recommend a relatively reasonable standing position by only providing a landscape image in portrait landscape photography, which greatly increases the availability.

INDEX TERMS Position recommendation, portrait generation, space transformer network.

I. INTRODUCTION

With the popularity of smartphones and digital cameras, people's enthusiasm for photography is increasing daily. The photographic works, especially the commemorative photos taken during traveling, are remarkable. When traveling to a certain place and seeing the beautiful scenery, people cannot help but take photos as souvenirs. Choosing a suitable standing position can help in taking beautiful photos. Selecting an appropriate standing position for a portrait to capture a photo that adheres to compositional aesthetics poses a captivating and intricate undertaking. This task relates to many practical, real-world applications, such as image generation,

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

editing, etc. The offline use of these technologies also has a very important application prospect.

The position relationship of foreground and background is related to the rationality of portrait and landscape photos. Therefore, where to place the portrait in the landscape image is an important task. In 2015, Wang et al. [1] proposed a reasonable portrait position recommendation method by weighing the scores of positive and negative rules. In this method, the positions where portraits are often located in group photos are set as positive rules, and the positions where portraits do not appear are negative. Finally, the final position recommendation is made according to the scores of positive and negative rules for arbitrary input scenes. In 2016, Zhang et al. [2] proposed recommending portrait positions through 3D aesthetic evaluation. In this method, a portrait foreground needs to be input in advance, and then the

foreground is traversed in the background image to generate multiple portrait photos. Each portrait photo is evaluated aesthetically, and the standing position of the portrait in the portrait photo with the highest score is taken as the optimal position. In 2018, Sheng et al. [3] developed a portrait photography navigation system to guide users to take images. In the method, scene images are used as input, professional photographic works with similar aesthetic characteristics are searched as examples, and the picturing relationship between examples is established using the random forest to recommend portrait positions. Although ordinary users can use the above method, recommend a suitable standing position to take a portrait landscape photo. Still, this method needs to input the portrait foreground image in advance. In recommending the portrait position, it needs prior knowledge to calculate the relevant position according to the scene image, or it does not consider the aesthetic features such as the composition and layout between the portrait foreground and the background. Due to the above disadvantages, the recommended portrait standing position has certain limitations. In the summary in Table 1, we list some research methods on portrait location recommendation and compare them with our proposed method.

As shown in Table 1, the existing research needs to input the foreground image of the portrait in advance, and the aesthetic features, such as the composition between the foreground and the background of the portrait, are not considered enough. The final recommendation results rarely generate real and complete photos. Therefore, we propose a position recommendation network (PositionRecNet) to solve the above problems based on landscape image and composition information. This network does not need to input portrait foreground images in advance. For the input landscape image, it can be recommended that portraits stand in proper positions in landscape images by considering the correlation between portrait foreground and landscape background and aesthetic characteristics such as composition. There is no uniform benchmark dataset in the existing portrait location recommendation research, which is also an important factor restricting the development of portrait location recommendation research [4]. To solve this problem, we have pre-built the People Landscape Dataset (PLDataset), which meets the task requirements. In the process of portrait location recommendation, this paper does not give a specific portrait foreground image but randomly generates the portrait foreground image through the pre-trained portrait generation network. First, a portrait generation network is trained based on the constructed dataset. Then, a spatial transformer network is used to change the position, size, and size of the portrait foreground image to fit the background image under the premise of inputting the landscape image. Finally, a composed network is applied to predict the position of the portrait in the landscape image and synthesize the portrait foreground and landscape background to generate the final result image. To sum up, the overall architecture of the portrait location recommendation model proposed in this paper

mainly involves the following modules: 1) Portrait generation network, which generates portrait foreground images through random input vectors. 2) Space transformer network STN. The generated portrait foreground and landscape background images are input into the STN network to generate a portrait foreground with an appropriate size and position that conforms to the background. 3) Composed network, which generates a realistic portrait landscape image. The converted portrait foreground and landscape background images are input into the compose module, and a proper standing position is recommended for the portrait. Finally, a more realistic portrait landscape image meeting the aesthetic characteristics of composition is generated. Experimental results on our datasets show that the proposed portrait position recommendation model can recommend proper standing positions for portraits and generate portrait landscapes that meet the composition rules. The contributions of this paper are as follows:

- 1) We propose a portrait position recommendation network named PositionRecNet. The network does not require a priori provision of the portrait foreground but solely utilizes unprompted landscape imagery as input, considering aesthetic elements like composition. Subsequently, it recommends an appropriate standing posture of the resulting portrait foreground.
- 2) We present the first Portrait Landscape Photo dataset (PLDataset) for portrait location recommendations.

TABLE 1. Portrait position recommendation methods.

Method	Deep learning	Aesthetic features	Without portrait foreground in advance	Results with real portrait
Our	*	*	*	*
Zhang et al. [5]		*		
Xu et al. [6]		*		
Wang et al. [1]		*	*	
Xu et al. [8]		*		
Tan et al. [9]	*	*		*
Lee et al. [10]	*			
Tripathi et al. [11]	*	*		*
Song et al. [24]	*	*		*

II. RELATED WORKS

A. TRADITIONAL METHODS FOR PORTRAIT LOCATION RECOMMENDATION

In the early stage of portrait position recommendation research, traditional methods were used to calculate the recommended position by combining prior knowledge and basic photography rules. In 2012, Zhang et al. [5] used the visual saliency model to extract attention synthesis features in professional photos. They proposed a geometric composite feature to learn spatial similarity to generate appropriate posture and position. In 2014, Xu et al. [6] studied the problem of

human position recommendation in mobile photography and proposed a human position recommendation strategy with a sensitive camera viewpoint. This strategy is based on the method of 3D reconstruction to align the background area and the human body area into a unified coordinate system. In the same year, Ma et al. [7] utilized an effective hierarchical search strategy to obtain the best position for portraits in a scene by researching professional photographs to determine a suitable portrait size and place people in images of a given scene. In 2015, Wang et al. [1] studied the positional relationship between portraits and scenes, defining negation rules to exclude unwanted components by combining the learned positive rules with the proposed negative rules to optimize the human subject's position in a given background scene. In the same year, Xu et al. [8] dynamically gave a given portrait standing by looking for an area of interest (ROI) from a given scene graph and then calculating to what extent the ROI conforms to the one-third dividing line and intersection rule, and whether its scale is close to one-third. In 2016, Zhang et al. [2] proposed a method to calculate layout recommendations to help with portrait positioning. First, 3D estimation generates several composite photos with different layouts. The resulting photos are then sequenced using a 3D layout aesthetic evaluation model. Finally, the layout features designed in the high-scoring photo are selected as the best layout and translated into a suitable position where the portrait should stand.

The above methods rely on traditional techniques. Although they potentially impact location recommendation, the computation in the recommendation process is very complex. It ignores semantics, composition, and other scene-specific information, which limits the final recommendation results.

B. DEEP LEARNING METHOD FOR PORTRAIT LOCATION RECOMMENDATION

Recently, deep learning methods have been applied to the research on portrait position recommendation. In 2018, Tan et al. [9] used a CNN network to predict the position and size of each potential portrait instance. In the same year, Lee et al. [10] proposed a network of two-generation modules based on a given semantic input map, one of which determines the proper insertion position of the object mask, and the other module determines the reasonable shape of the object mask. In 2019, Tripathi et al. [11] proposed a three-way competition between the generative network, the target network, and the discriminator for image authenticity. Given a background and foreground mask, the generative network generates a composite image by learning affine transformations. Then, the target network detects the position of all foregrounds in the composite image to determine the plausibility of the position. In 2020, Zhang et al. [12] combined the encoded object and background with a random variable to predict the object's position and then a discriminator that discriminates based on foreground and background to judge the plausibility of the position. In addition, the diversity of object

layouts is maintained by predicting the pairwise distance between the layout and the corresponding random variable. In 2022, Song et al. [24] designed a composition recommendation model based on pose attributes to learn composition rules. They believe that the composition of portrait photos is related to the scene's visual content and the human body's posture. Roy et al. [25] predict the new person's potential location and skeletal structure by conditioning a Wasserstein Generative Adversarial Network (WGAN) on the existing human skeletons in the scene.

Although the above method adopts deep learning methods, they require a pre-specified portrait or the anchor frame position information of the portrait in the scene. They do not consider aesthetic features such as the composition between the recommended portrait and the background.

C. IMAGE COMPOSITION

Image composition aims to cut the foreground from one image and paste it on another, resulting in a composite image [26], [27]. Several issues can cause composite images to be unrealistic. For example, incompatible lighting [28], the unreasonable relative size of foreground and background [29], semantic mismatch of foreground and background [30], etc.

D. GAN-BASED IMAGE GENERATION

Generative Adversarial Network (GAN) is a powerful deep learning model [31]. It has many types [32] [33], and it has also achieved remarkable development in many fields, such as image synthesis [34]. Some recent GAN models, such as StyleGAN2 [13], demonstrate the powerful ability of GANs to generate images that are almost indistinguishable from real images.

III. METHODOLOGY

A. THE PORTRAIT POSITION RECOMMENDATION MODEL

When we want to record a photo of ourselves and the surrounding beauty in reality, how the photographer should compose the photo to take a better photo becomes a difficult problem. This paper presents a position recommendation model framework, as shown in Figure 1. This model is based on an input pure landscape image, which can recommend the portrait's appropriate standing position in the landscape image and, finally, output a portrait landscape image that conforms to the aesthetic characteristics of the composition. The specific steps are as follows:

Firstly, we need to determine the foreground image of a human image to represent the size of a human image's position. For a certain scenery, there will be different people taking images here. Exclusively relying on a particular portrait foreground as input may not align with the objective reality of diverse individuals capturing images in this context. Therefore, considering the objective facts of different people's portrait prospects, this paper does not use portrait prospects as the input of the location recommendation

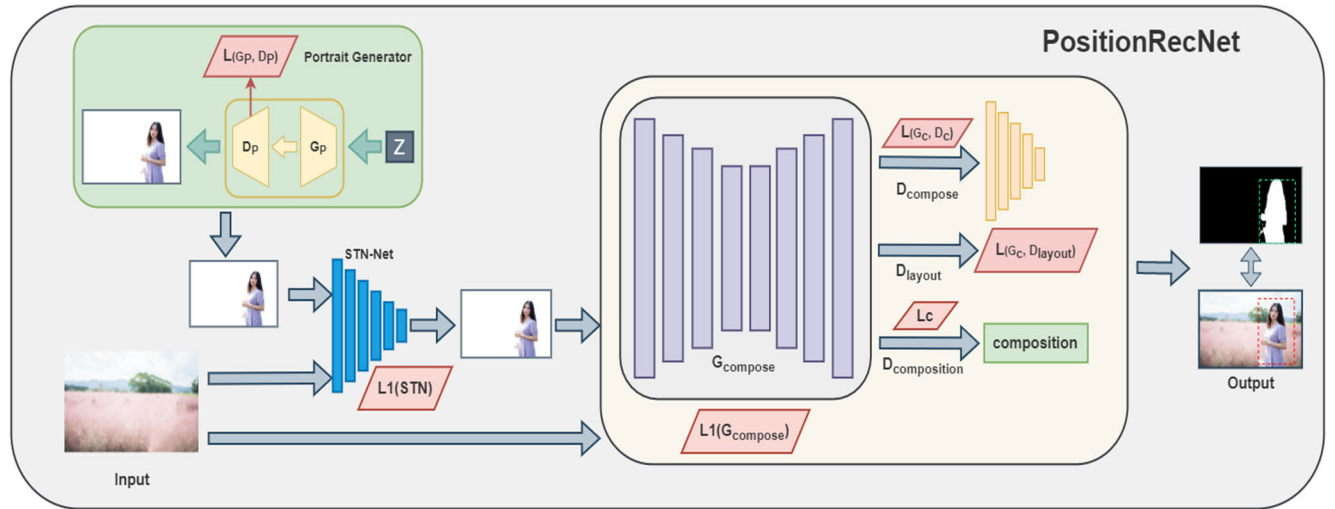


FIGURE 1. The position recommendation network (PositionRecNet) includes a portrait generation network, space transformer network STN, and composed network.

network. We propose a portrait generation network to obtain portrait foreground images of any size without a specific portrait input, as shown in the green box in Figure 1. The network uses the random vector Z to generate a random portrait foreground image.

Secondly, leveraging randomly selected landscape images as inputs to PositionRecNet, in conjunction with the previously generated randomly synthesized portrait images, we propose recommendations for appropriate standing positions within the landscape images for the corresponding portrait images. In order to make the portrait foreground and the landscape background not only conform to the rationality of the position but also consider the basic aesthetic features, such as the correlation between the front and rear scenes and composition rules, we use the space transformation network STN to recommend a suitable standing position for portraits according to the landscape.

Finally, the landscape background image and the transformed portrait foreground image are fed to the Compose module. Our research aims to recommend a suitable standing position for portraits so that users can perceive the result more intuitively and clearly when the final recommendation result is a relatively real picture. Therefore, we introduce the Compose module to synthesize relatively real recommendation results. This module not only considers the correlation between portrait foreground and landscape background but also incorporates the overall aesthetics of the composition. It effectively suggests the best standing positions of portraits against landscape backgrounds, resulting in more realistic portrait landscape images. Table 2 lists the relevant symbols used in this paper and their corresponding meanings.

This paper adopts StyleGAN [13], [14], [15] for the portrait generation network. Based on the random noise Z input randomly, the purpose of the portrait foreground generation network is to generate the portrait foreground matching the background size according to the random noise z . That is $G_p(z) = P$.

In our model, the portrait foreground image needs to be scaled and spatially changed based on the input landscape background image to make it fit with the landscape background image. It also involves the problem of determining the composition class of the whole image. To make the generated portrait foreground image fit with the input landscape background image, a spatial transformer network STN is introduced in this paper, which expresses the position and size changes of portraits in the landscape background image in the form of affine changes. The specific principle is shown in Formula 1:

$$(x', y') = \begin{pmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{pmatrix} (x, y), \quad (1)$$

Which (x, y) represents the position coordinates of a pixel in the image and (x', y') represents the position coordinates of the pixel after the affine transformation matrix. θ_{13}, θ_{23} indicate a translation operation. $\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}$ represent linear operations such as rotation and scaling.

For the compose module, this paper adopts the idea of GANs [16], [17]. In the compose module, it is necessary to consider the correlation between portrait foreground and landscape background and the aesthetic characteristics, such as the overall image composition. We introduce a generator to recommend the proper portrait position in the landscape to achieve this goal, considering factors such as composition aesthetics between the front and back scenes. $G_{compose}$ is a good standing position for portraits recommended in landscape images under the premise of considering factors such as the composition aesthetics between the front and back scenes. We mainly judge the image generation effect and the rationality of the recommended portrait position in the process of portrait position recommendation. For the rationality of the standing portrait position, this paper mainly determines the rationality of the recommended position by considering the correlation between generated and real portrait masks, as shown in Figure 2. Specifically, the corresponding mask

TABLE 2. Symbols and meanings used in this paper.

symbol	meaning	symbol	meaning
G_p	Generate a portrait foreground image	D_p	Discriminatively generated portrait foreground image
$G_{compose}$	Generate portrait landscape image	$D_{compose}$	Discriminatively generated portrait landscape image
$L_{(G_p, D_p)}$	Loss of authenticity of portrait foreground image generation	$L_{(G_c, D_c)}$	Loss of authenticity of portrait landscape image generation
$L1(G_{compose})$	Pixel loss between portrait landscape generated image and original image	$L1(STN)$	Image pixel loss before and after spatial network conversion
L^{recon}	Pixel reconstruction loss	L_c	Cross-entropy loss of composition between generated landscape image and original image
$L_{(G_c, D_{layout})}$	Loss of portrait position in portrait landscape		

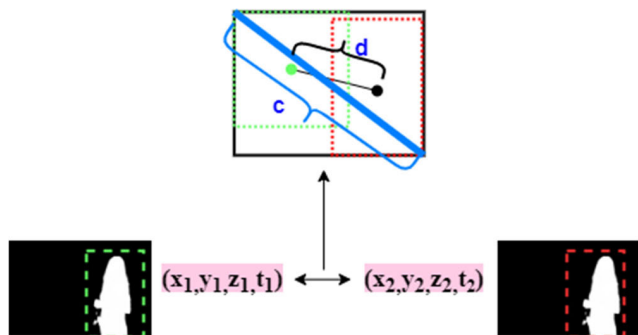


FIGURE 2. Clou in portrait standing position.

image is obtained by the generated portrait landscape image, and the mask image obtains the central coordinates of the portrait. Then, the central coordinates calibrate the anchor frame position coordinate information of the portrait in the image. Finally, the judgment is made by the anchor frame coordinates of the portrait in the generated image and the anchor frame information of the portrait in the real image.

Since the pre-trained portrait generating network randomly generates the foreground image of the portrait according to the random input vector, we cannot guarantee the complete matching degree of the generated portrait prospect image and the real portrait foreground image. Therefore, we choose not to directly determine the plausibility of the portrait's proposed standing position by computing the Clou index between the mask of the generated image and the real image. Our focus primarily centers on the portrait's standing position, thereby emphasizing the significance of coordinate information, including the central coordinates of the portrait and the fundamental anchor frame coordinates within the image. As for the aesthetic characteristics of the overall image composition, this paper mainly makes a more intuitive judgment on the composition class between the generated and natural landscapes.

B. LOSS FUNCTION

For the portrait generation network, this paper mainly adopts the idea of StyleGAN. That is, the confrontation loss shown in Formula 4 is set between the portrait generator G_p and the

portrait discriminator D_p .

$$Loss_G = \log(\exp(-D(G(z))) + 1), \quad (2)$$

$$Loss_D = \log(\exp(D(G(z))) + 1) + \log(\exp(-D(x)) + 1), \quad (3)$$

$$L(G_p, D_p) = Loss_G + Loss_D, \quad (4)$$

Among them, z represents a random vector.

Based on the above pre-conditions, such as image authenticity, the rationality of portrait position, and aesthetic characteristics of image composition, this paper conducts relevant experiments through the overall loss function shown in Formula 5.

$$L = \alpha L_{(G_c, D_c)} + \beta L_{(G_c, D_{layout})} + \gamma L^{recon} + \delta L_c, \quad (5)$$

The first part is the loss of portrait landscape image generation. The second part is the rationality loss of portrait position recommendation; The third part is the reconstruction loss of pixels; The fourth part is the cross entropy loss of composition classes of portrait landscape as a whole.

1) LOSS OF PORTRAIT LANDSCAPE IMAGE GENERATION

The loss of authenticity of portrait landscape image generation is shown in Formula 6:

$$L_{(G_c, D_c)} = E_{(F, B, C)} \left[\log D_{compose}(F', B', C^T) \right] + E_{(F, B)} \left[1 - \log D_{compose}(F', B', C) \right], \quad (6)$$

where F represents the foreground image, B represents the background image, C represents the generated portrait landscape image, C^T represents the real portrait landscape image. As shown in Formula 7, the foreground image and background image can be input into the space transformation model STN to obtain the affine transformed foreground image F' and background image B' .

$$(F', B') = STN(F, B), \quad (7)$$

2) THE RATIONALITY LOSS OF PORTRAIT POSITION RECOMMENDATION

The position loss of generated portrait foreground in the landscape is $L_{(G_c, D_{layout})}$ as shown in Equation 8. For the

position loss of the generated portrait in the landscape, this paper uses CIoU loss [18] to make a measurement.

$$L_{(G_c, D_{layout})} = CIoU = IoU - \frac{\rho^2(b, b^{gt})}{c^2} - \alpha v, \quad (8)$$

Among them, the Euclidean distance between the predicted position and the center point of the real anchor frame and the diagonal distance that can simultaneously contain the smallest closed area between the predicted and real anchor frames are shown in Formulas 9 and 10, respectively. $\rho^2(b, b^{gt})$ indicates the Euclidean distance between the predicted position and the center point of the real position anchor box, c indicates the distance between diagonals that can contain the smallest closed area between the predicted anchor box and the real anchor box, the formulas of α and v are as follows:

$$\alpha = \frac{v}{1 - IoU + v}, \quad (9)$$

$$v = \frac{4}{\pi^2} \left(\arctan \frac{w^{gt}}{h^{gt}} - \arctan \frac{w}{h} \right)^2, \quad (10)$$

Compared with IoU , the position information between the predicted anchor frame and the real anchor frame is considered, which solves the problem that the loss is zero if there is no overlap between them. Adding the length and width information of the anchor frame can make the prediction more realistic and accelerate the convergence.

3) PIXEL RECONSTRUCTION LOSS

Although antagonistic loss aims to establish the distribution model of objects in the data, it is often prone to model collapse because it does not cover the distribution of the whole mode [17]. Therefore, this paper adds the reconstruction loss L^{recon} between pixels as shown in Formula 11:

$$L^{recon} = L1(G_{compose}) + L1(STN), \quad (11)$$

$$L1(G_{compose}) = E_{(F, B, C)} \left[\left\| C - C^T \right\|_1 \right], \quad (12)$$

$$L1(STN) = E_{(F, B)} \left[\left\| (F^T, B^T) - (F', B') \right\|_1 \right], \quad (13)$$

Among them, the real portrait foreground image and landscape background image are respectively represented as F^T, B^T .

C. CONSTRUCTION OF DATASET

Currently, for the portrait position recommendation task, the dataset used by the existing methods is roughly constructed by users randomly crawling on the Internet according to their own needs and then manually selecting. Hence, there is no benchmark universal dataset available. Portrait position recommendation task generally requires a group photo with aesthetic characteristics such as composition, but randomly selected images cannot guarantee image quality. This paper mainly focuses on recommending a suitable position for portraits that meet the aesthetic characteristics of a composition according to a given landscape image. To solve this problem, we build our portrait recommendation dataset based on the



FIGURE 3. Portrait dataset examples.

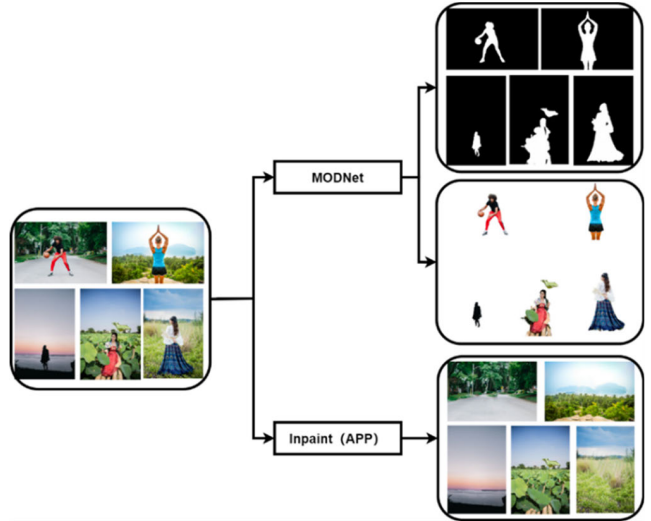


FIGURE 4. Workflow for generating image pairs used in network training. Selected portrait landscape images with aesthetic composition features are processed using MODNet and Inpaint software to obtain portrait subject masks and pure landscape images without human subjects.

PPR10K dataset [19]. Because professional photographers take the PPR10K dataset, the portrait images of the PPR10K can be considered to conform to the aesthetic composition's characteristics. We screened 3024 portrait and landscape photos to construct the required dataset. As shown in Figure 4, to obtain image pairs for network training, we output the selected portrait and landscape images with compositional aesthetics to the MODNet network and Inpaint software for processing [20]. In the end, we obtained the mask of the portrait landscape and the pure landscape without the portrait subject.

To obtain a better portrait generation effect, we use the DeepFashion dataset [21], [22], 6415 pure portrait images were selected from CCP to train the portrait generation network. The selected portrait images are roughly shown in Figure 3.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

A. IMPLEMENTATION DETAILS (MODEL FRAMEWORK, PARAMETER SETTING, EVALUATION CRITERIA)

Our model is implemented by Tensorflow architecture, and all experiments are conducted on Tesla V100-DGXS GPU.

The frame is fine-tuned based on StyleGAN in this paper on portrait generation networks. $C(k)$ indicates that $K \times 4 \times 4$ convolution kernels are used in this convolution layer

with a step distance 2. L (n) indicates the full connection layer that outputs N-dimensional parameters. In the character location recommendation network in this paper, the framework of the generator is C (32)-C (64)-C (128)-C (256)-C (512)-L (256)-L (8), and the framework of the discriminator is C (32)-C (64)-C (11). The ReLU activation function is used for each layer of the generator, and the LeakyReLU function with a parameter of 0.2 is used for each layer of the discriminator.

Setting parameters During the experiment in this paper, all models were trained for 2000 rounds, and the training started with a learning rate of = 0.0001, and the batch size was set to 20. Loss part, setting, $\epsilon = 0.0001, \alpha = \beta = \gamma = \delta = 1$.

To verify the method's effectiveness in this paper, we measure the structural similarity of SSIM and CIoU indicators in the quantitative analysis part.

1) EVALUATION CRITERIA

To measure the similarity between the generated and real images, this paper uses the structural similarity loss function SSIM [23]. The structure similarity function mainly considers the image's brightness, contrast, and structure, as shown in Formula 14.

$$SSIM = \frac{(2\mu_x\mu_y + \epsilon_1)(2\sigma_{xy} + \epsilon_2)}{(\mu_x^2 + \mu_y^2 + \epsilon_1)(\sigma_x^2 + \sigma_y^2 + \epsilon_2)}, \quad (14)$$

where two images are respectively represented x, y , and the mean values of the images are respectively represented μ_x, μ_y ; The variance and covariance of σ_x, σ_y and σ_{xy} , respectively, are two stable variables. ϵ_1 and ϵ_2 are two variables that maintain stability, as shown in Formula 15. Among them, l is the dynamic range of pixels. Generally, it is the default $k_1 = 0.01, k_2 = 0.03$.

$$\epsilon_1 = (k_1l)^2, \epsilon_2 = (k_2l)^2, \quad (15)$$

B. THE RESULTS OF THE PORTRAIT GENERATION NETWORK

The portrait generation model proposed in this paper can generate a relatively real portrait foreground image through random noise input at random, to provide a precursor for the location recommendation network.

Figure 5 shows the prediction results of the portrait generation network proposed in this paper. It can be seen from the figure that compared with the original GAN and DCGAN, the portrait generation network used in this paper has relatively few artifacts and distortions. The purpose of the portrait generation network in this paper is to generate a rough portrait shape to indicate the position and size of the portrait, and its authenticity is relatively limited.

C. THE EXPERIMENTAL RESULTS OF PORTRAIT POSITION RECOMMENDATION

The main purpose of our portrait position recommendation network is to hope the network can recommend suitable standing positions for people according to the input



FIGURE 5. Prediction results of portrait generation network.



FIGURE 6. The results of the portrait position recommendation. Among them, the first, third, and fifth rows are the results of location recommendation and people-scene photos of this paper (OURS), and the real results (GT) correspond to the second, fourth, and sixth rows.

landscape image. The portrait position recommendation results are shown in Figure 6.

As can be seen from Figure 6, the proposed PositionRecNet can recommend randomly generated portrait foreground images to proper standing positions in landscape images according to the input landscape images. The effect of the portrait foreground limits the generated portrait group photo, and its authenticity is relatively biased. Still, the effectiveness of the position recommendation effect in this paper cannot be denied.

We also use the proposed method to recommend corresponding proper positions for portraits for common scenarios such as roads, outdoors, seaside, and buildings. The recommended results are shown in Figure 7.

The portrait foreground image generated by a random vector has some defects. Therefore, to further verify the effect of the portrait position recommendation model proposed in this paper, we take the real portrait foreground image as input to explore further the effect of the proposed method in this paper.

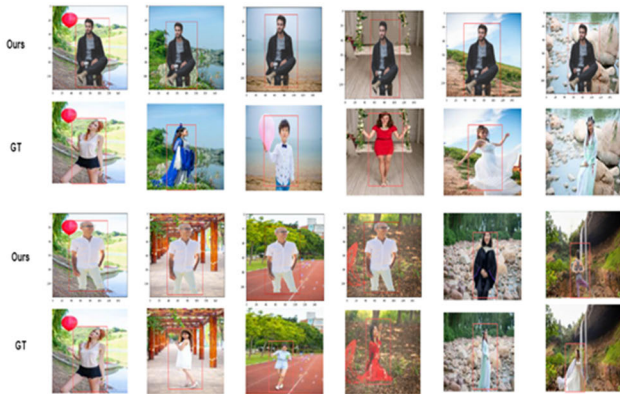


FIGURE 7. Location recommendation for different scenarios.

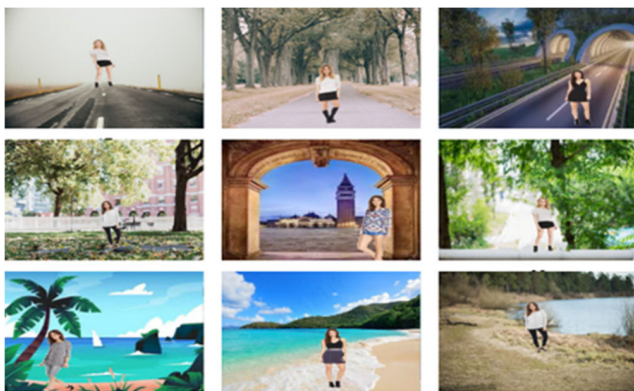


FIGURE 8. Recommend results based on the location of a given specific portrait. Ours represents the effect generated by the method in this paper, and GT represents a real portrait photo.

As shown in Figure 8, given the specific portrait foreground image, our method can better recommend suitable standing positions in input landscape background images. From the first and third rows of Figure 7, it can be seen that the portrait position recommendation model proposed in this paper can recommend the same portrait to a proper position in different landscapes; As can be seen from the images in the first column, our portrait position recommendation model can also recommend different portraits to good standing positions in the same landscape. The above results can better reflect the rationality of the portrait position recommendation model proposed in this paper.

The indicators of average structural similarity SSIM and CIoU evaluate the proposed method quantitatively. The idea of average structural similarity SSIM is to divide images into blocks by sliding windows so that the total number of blocks is N . Considering the influence of window shapes on blocks, each window's average, variance, and covariance are calculated by Gaussian weighting. Then the structural similarity SSIM of the corresponding blocks is calculated. Finally, the average is used as the structural similarity measure of the two images. SSIM is a number between 0 and 1. The larger the number, the smaller the gap between the output and undistorted images. That is, the better the image quality. CIoU

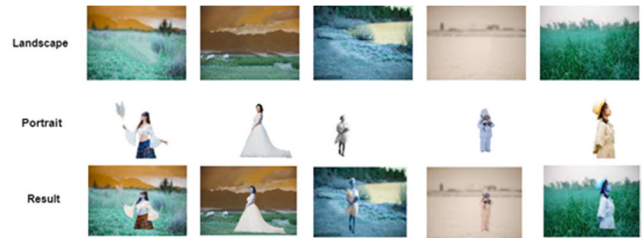


FIGURE 9. The results of STN.

index is used to judge the matching degree between the recommended portrait standing and real portrait positions. The larger the value of CIoU, the more accurate the recommended position is.

Through quantitative experimental analysis, the structural similarity SSIM and CIoU between the portrait landscape finally generated by the location recommendation method proposed in this paper and the real portrait landscape are 0.69014 and 0.61568, respectively. From the results, the method in this paper has a good matching degree, which reflects the rationality of the method from a quantitative point of view.

D. THE RESULT OF THE STN MODEL

In the portrait position recommendation model, to make the generated portrait foreground image match the landscape background image according to the input background image, this paper introduces the space transformation network STN. This section, as shown in Figure 9, mainly shows the result of transforming the portrait by the spatial transformer network according to the input background.








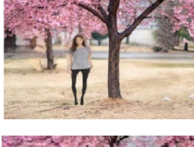

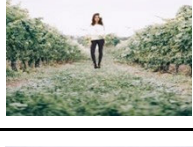

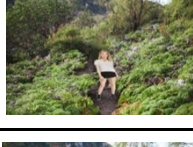


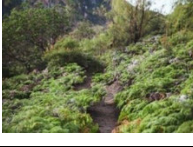



As evident from the illustrated figure, the Spatial Transformer Network (STN) employed in this research effectively incorporates pertinent features from the landscape background image, thereby enhancing the alignment between the portrait foreground and the landscape background. This alignment lays the groundwork and ensures the subsequent location recommendation process.

E. ABLATION EXPERIMENTS

In the methods of this paper, we mainly use the loss of portrait style drawing generation (Basic Loss), composition class loss (L_c), portrait position determination loss ($L_{(G_c, D_{layout})}$), and image reconstruction loss (L_{recon}). In this ablation experiment part, we mainly discuss the influence of various loss functions on the experimental results. We discussed whether not using composition class loss, portrait position determination loss, and image reconstruction loss have a certain impact on the final recommended portrait position determination effect.

It can be seen from the ablation experiment in Table 3 that if the loss of image reconstruction is not added, as shown in the first row of the table, the generated image is difficult to present if the loss of composition class or the loss of position determination leads to a greater deviation in the final position

TABLE 3. Ablation experiment.

Loss Function				Result		
Basic	L^{recon}	L_c	$L_{(G_c, D_{layout})}$			
✓		✓	✓			
✓	✓		✓			
✓	✓	✓				
✓	✓	✓	✓			
Background image						
Ground Truth						

recommendation result. If all the above losses are added in the experimental process of this paper, the final position recommendation is more reasonable. The generated portrait landscape is relatively better, thus reflecting the effect of the portrait position recommendation model in this paper.

F. QUALITATIVE RESEARCH

1) QUALITATIVE EVALUATION VIA USER STUDY

Users have different aesthetic preferences, but we can only recommend the best location due to computer limitations. Therefore, we introduce a user study to qualitatively judge the feasibility of the method proposed in this paper.

We randomly selected 50 generated portrait landscape images from the test results to form the user evaluation dataset in this paper. We invited 60 people from different professional and educational backgrounds, including 30 girls and 30 boys, to rate 50 images through a questionnaire. We set five scores for each image: 1, 2, 3, 4, and 5, and we evaluate

the rationality of our method by calculating the arithmetic mean of each portrait group photo. When the average score of each image is greater than 3.5 points, it indicates that the recommended position is consistent with the user’s expected portrait standing position; the average score of each image is between 2 and 3.5 points, indicating that the recommended position is slightly different from the user’s idea; The average score per image is less than 2 points, which means that the recommended location is very different from the user’s desired location.

Figures 10 and 11 show the average score and the overall plausibility score distribution for these 50 portrait photos. As shown in Figure 11, 60% of people think the recommended location matches the user’s desired location, 28% think the suggested location is relatively reasonable, and 12% think it is unreasonable. According to the statistical results, we believe that the portrait position recommendation method proposed in this paper meets the cognitive needs of the public and has a certain recommendation effect.

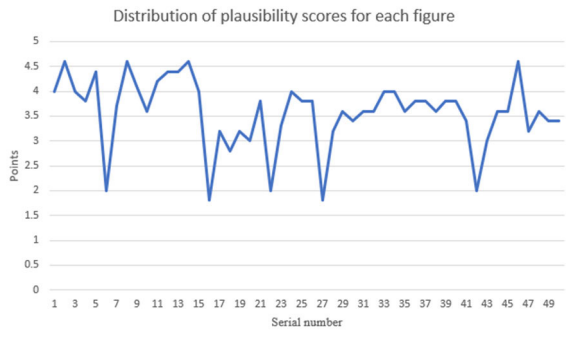


FIGURE 10. Score display for each portrait group photo.

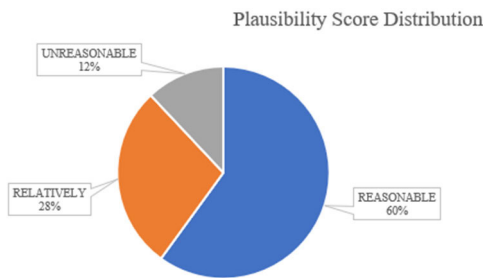


FIGURE 11. Rationality score distribution of portrait position recommendation.

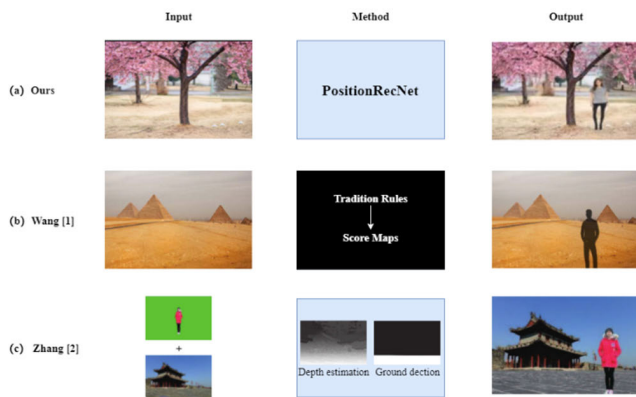


FIGURE 12. Comparison of our method with state-of-the-art models in the field of portrait position recommendation.

2) QUALITATIVE COMPARISON WITH STATE-OF-THE-ART MODELS

For the portrait location recommendation domain, we compare our recommendation process and results with state-of-the-art models, as shown in Figure 12. Compared with Wang [1], our model is an end-to-end recommendation process, and the recommendation results we provide are real images instead of masks. Our results are presented more intuitively and clearly. Compared with Zhang [2], our model does not need to provide the input of the portrait foreground in advance, which is more in line with people’s daily shooting situations.

G. LIMITATION ANALYSIS

The main purpose of this paper is to recommend a suitable standing position for portraits. For the final portrait photo, the image’s authenticity is relatively poor. The main reason lies in the part of the portrait generation network. Considering that the same kind of scene may correspond to different portraits, this paper considers the landscape as the only input in the portrait position recommendation model. The pre-trained portrait generation model generates a portrait foreground image through random noise. In generating the portrait foreground image, the authenticity of the portrait foreground image is not considered too much, and the generated portrait does not consider the landscape image’s characteristics. The size of the portrait is re-obtained, so the foreground image of the generated portrait has relatively poor realism, and there are also problems such as deformation. The coordination and clarity of the currently produced portrait and landscape images are also insufficient. However, the application scenario of our proposed method is mainly real-time shooting guidance, and its focus is on recommending portrait locations, so the deformation and unclearness of the generated results do not affect the final recommendation results. We consider adding recommended person poses [35], [36] and using clearer images to deal with these issues in the future.

V. CONCLUSION AND FUTURE WORKS

This paper mainly considers where the portrait should stand when people want to take a portrait landscape photo when they see a beautiful landscape. Aiming at this practical problem, this paper puts forward an end-to-end portrait position recommendation model, which can recommend the good standing position of the portrait in the landscape image for the input landscape image and finally generate a relatively real portrait landscape image. Portrait position recommendation mainly comprises a portrait foreground generation network, spatial transformer network, and composed network. According to the task’s requirements, this paper constructs the portrait landscape dataset PLDataset for the corresponding training of the model. The experimental results on this dataset show that the portrait position recommendation model proposed in this paper can recommend a relatively proper standing position for portraits. The research results of this paper can be used for practical applications and later tasks such as offline image processing. Aiming at the problem that the generated portrait landscape images are not authentic enough, our future research considers the portrait’s position and combines the semantic information of the background further to enhance the authenticity of the generated portrait foreground images, and we can also consider adding pose information.

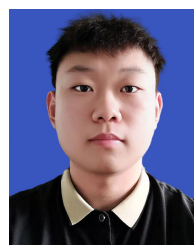
REFERENCES

[1] Y. Wang, M. Song, D. Tao, Y. Rui, J. Bu, A. C. Tsoi, S. Zhuo, and P. Tan, “Where2Stand: A human position recommendation system for souvenir photography,” *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 1, pp. 1–22, Oct. 2015.

- [2] B. Zhang, R. Ju, T. Ren, and G. Wu, "Say cheese: Personal photography layout recommendation using 3-D aesthetics estimation," in *Proc. Pacific Rim Conf. Multimedia*. Cham, Switzerland: Springer, 2016, pp. 13–23.
- [3] B. Sheng, Y. Jin, P. Li, W. Wang, H. Fu, and E. Wu, "InspireMePosing: Learn pose and composition from portrait examples," in *Proc. PG (Short Papers Posters)*, 2018, pp. 33–35.
- [4] K. Kafle and C. Kanan, "Visual question answering: Datasets, algorithms, and future challenges," *Comput. Vis. Image Understand.*, vol. 163, pp. 3–20, Oct. 2017.
- [5] Y. Zhang, X. Sun, H. Yao, L. Qin, and Q. Huang, "Aesthetic composition representation for portrait photographing recommendation," in *Proc. 19th IEEE Int. Conf. Image Process.*, Sep. 2012, pp. 2753–2756.
- [6] S. Ma, Y. Fan, and C. W. Chen, "Finding your spot: A photography suggestion system for placing human in the scene," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 556–560.
- [7] S. Ma, Y. Fan, and C. W. Chen, "Pose maker: A pose recommendation system for person in the landscape photographing," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 1053–1056.
- [8] Y. Xu, J. Ratcliff, J. Scovell, G. Speiginer, and R. Azuma, "Real-time guidance camera interface to enhance photo aesthetic quality," in *Proc. 33rd Annu. ACM Conf. Human Factors Comput. Syst.*, 2015, pp. 1183–1186.
- [9] F. Tan, C. Bernier, B. Cohen, V. Ordonez, and C. Barnes, "Where and who? Automatic semantic-aware person composition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1519–1528.
- [10] Y. S. Rawat and M. S. Kankanhalli, "Context-aware photography learning for smart mobile devices," *ACM Trans. Multimedia Comput., Commun., Appl. (TOMM)*, vol. 12, pp. 1–24, Oct. 2015.
- [11] S. Tripathi, S. Chandra, A. Agrawal, A. Tyagi, J. M. Rehg, and V. Chari, "Learning to generate synthetic data via compositing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 461–470.
- [12] L. Zhang, T. Wen, J. Min, J. Wang, D. Han, and J. Shi, "Learning object placement by inpainting for compositional data augmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 566–581.
- [13] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8110–8119.
- [14] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 852–863.
- [15] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4396–4405.
- [16] J.-Y. 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. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.
- [17] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.
- [18] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
- [19] J. Liang, H. Zeng, M. Cui, X. Xie, and L. Zhang, "PPR10K: A large-scale portrait photo retouching dataset with human-region mask and group-level consistency," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 653–661.
- [20] Z. Ke, J. Sun, K. Li, Q. Yan, and R. W. Lau, "ModNet: Real-time trimap-free portrait matting via objective decomposition," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 1, pp. 1140–1147.
- [21] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1096–1104.
- [22] Y. Ge, R. Zhang, X. Wang, X. Tang, and P. Luo, "DeepFashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5332–5340.
- [23] Z. Wang, "Image quality assessment: From error measurement to structural similarity," *IEEE Trans. Image Process.*, vol. 13, pp. 600–613, 2003.
- [24] X. Song, J. Pan, F. Wu, and W. Dong, "Optimal composition recommendation for portrait photography," in *Proc. SIGGRAPH Asia Posters*, Dec. 2022, pp. 1–2.
- [25] P. Roy, S. Ghosh, S. Bhattacharya, U. Pal, and M. Blumenstein, "Scene aware person image generation through global contextual conditioning," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 2764–2770.
- [26] L. Niu, W. Cong, L. Liu, Y. Hong, B. Zhang, J. Liang, and L. Zhang, "Making images real again: A comprehensive survey on deep image composition," 2021, *arXiv:2106.14490*.
- [27] C.-H. Lin, E. Yumer, O. Wang, E. Shechtman, and S. Lucey, "ST-GAN: Spatial transformer generative adversarial networks for image compositing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9455–9464.
- [28] R. Pandey, S. O. Escolano, C. Legendre, C. Häne, S. Bouaziz, C. Rhemann, P. Debevec, and S. Fanello, "Total relighting: Learning to relight portraits for background replacement," *ACM Trans. Graph.*, vol. 40, no. 4, pp. 1–21, Aug. 2021.
- [29] A. Volokitin, I. Susmelj, E. Agustsson, L. Van Gool, and R. Timofte, "Efficiently detecting plausible locations for object placement using masked convolutions," *Computer Vision—ECCV*, vol. 16. Glasgow, U.K.: Springer, 2020, pp. 252–266.
- [30] F. Zhan, H. Zhu, and S. Lu, "Spatial fusion GAN for image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3648–3657.
- [31] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.
- [32] G. Wei, J. Guo, Y. Ke, K. Wang, S. Yang, and N. Sheng, "A three-stage GAN model based on edge and color prediction for image outpainting," *Exp. Syst. Appl.*, vol. 214, Mar. 2023, Art. no. 119136.
- [33] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5907–5915.
- [34] P. Zhu, R. Abdal, J. Femiani, and P. Wonka, "Barbershop: GAN-based image compositing using segmentation masks," 2021, *arXiv:2106.01505*.
- [35] Y. Wang, S. Hou, B. Ning, and W. Liang, "Photo stand-out: Photography with virtual character," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 781–788.
- [36] T. Zhang, J. Lian, J. Wen, and C. L. P. Chen, "Multi-person pose estimation in the wild: Using adversarial method to train a top-down pose estimation network," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 53, no. 7, pp. 3919–3929, Jul. 2023.



ZHENQUAN SHI received the B.E. and M.E. degrees in computer application from Tiangong University, China, in 2001 and 2004, respectively. Since 2005, he has been with Tiangong University, where he is currently a Lecturer. His current research interests include the Internet of Things, image processing, deep learning, social networks, and information visualization.



QINGGANG HOU received the bachelor's degree from the Internet of Things engineering from Shandong Agriculture and Engineering University, China, in 2020, and the master's degree from the School of Computer Science and Technology, Tiangong University, China, in 2023.



GUANJUN SHENG received the B.E. degree in computer science and technology from the Tianjin University of Science and Technology, China, in 2022. He is currently pursuing the master's degree in computer science and technology with the School of Computer Science and Technology, Tiangong University, Tianjin, China.



KAI WANG received the B.E. degree in electronic information science and technology from Qingdao University, China, in 2010, and the M.E. degree in software engineering from Tiangong University, China, in 2017, on the research subjects of artificial intelligence and image processing. Since 2017, he has been with Tiangong University, where he is currently a Researcher with the School of Computer Science and Technology. His current research interests include image processing and information visualization.



YONGZHEN KE received the B.E. degree in computer application from Tianjin Polytechnic University, China, in 1997, and the M.E. and Ph.D. degrees in computer application from Tianjin University, China, in 2000 and 2008, respectively, on the research subjects of image processing and data visualization. Since 1997, he has been with Tiangong University, where he is currently a Professor with the School of Computer Science and Technology. His current research interests include image processing, intelligent image analysis, computational aesthetic, digital image forensic, and information visualization.



YUNGANG JIA received the bachelor's degree in computer software from the Tianjin University of Technology, in 2000, and the Master of Engineering degree in software engineering from Nankai University, in 2005. Since 2000, he has been with the Tianjin Branch of the National Computer Network Emergency Technology Coordination Center, China, where he is currently a Senior Engineer. His current research interests include network security, artificial intelligence, image processing, information visualization, and regional cross chain.

...