IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 27 May 2023, accepted 21 June 2023, date of publication 26 June 2023, date of current version 12 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3289590

RESEARCH ARTICLE

SCRN: Stepwise Change and Refine Network Based Semantic Distribution for Human Pose Transfer

HAN MO[1], YANG XU[1,2], YOUJU PENG[1], AND GUIDONG XU[1]
[1]College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China
[2]Guiyang Aluminum-Magnesium Design and Research Institute Company Ltd., Guiyang 550009, China

Corresponding author: Yang Xu (xuy@gzu.edu.cn)

This work was supported by the Guizhou Provincial Key Technology Research and Development Program [2023] General 326.

ABSTRACT It is a challenging and meaningful task to achieve person image synthesis by guiding pose. However, two problems have existed in past work: inaccurate generated poses and inconsistency with the target texture. To address these issues, we propose the Stepwise Change and Refine Network (SCRN), a two-stage network that aims to transfer given person images to the target pose while generating more reasonable and closer-to-real results. In the first stage, coarse images are generated using a series of modules with the same structure called Coarse Blocks. This process gradually changes the pose to achieve better shape consistency with the target image. In the second stage, style features are extracted from the original image by distributing semantic information. These features are used to optimize the rough image to obtain the final generated image, resulting in better consistency with the appearance of the target image. Our proposed method preserves both the pose's spatial features and the original image's texture features. Furthermore, we introduce a new loss function to make the generated image more in line with human perception. Qualitative and quantitative experiments with state-of-the-art models demonstrate significant improvements in SSIM, FID, PSNR, and LPIPS, validating the superiority of our model.

INDEX TERMS Deep learning, generative adversarial network, human pose transfer, skeleton based approach, person image synthesis.

I. INTRODUCTION

Person image synthesis is an important and challenging task in computer vision, with many practical applications such as virtual try-on [1] and video generation [2]. This paper proposes the Stepwise Change and Refine Network (SCRN) for person image synthesis through human pose transfer. As people move and engage in physical activities, their postures change frequently, making it important to synthesize images that accurately represent the desired pose. Figure 1 provides a classic example of pose transfer. As shown in Figure 1, our model takes the source image and different poses as inputs and generates an image that retains the source image portrait while incorporating the given pose.

The associate editor coordinating the review of this manuscript and approving it for publication was Eduardo Rosa-Molinar.

In pose transfer, it is necessary to analyze and extract the keypoint information of the source pose. This keypoint information is then mapped onto the target pose. Although pose transfer technology has been widely studied and applied in the field of computer vision, there are still many challenges and problems. For example, keypoints may be missing or mismatched during the pose transfer process. Huang et al. [3] proposes Pose Attentional Transfer Network (PATN) to guide the pose gradually by using the attention mechanism, and limit the change of pose in a small range in each transfer step, which can better preserve the spatial relationship and simplify the network structure. Although it has shown decent ability in this area, it lacks the ability to align source style and target pose. During the gradual transfer process, the style features of the source image may be gradually lost, resulting in loss of detail in the generated portrait's clothing

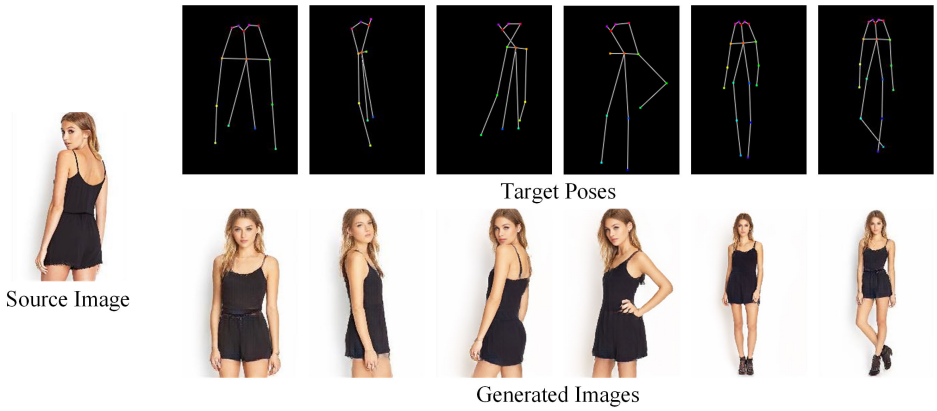


FIGURE 1. Different portrait poses generated by our model.

patterns and materials. Therefore, the module dedicated to extracting reference style features has been introduced into some subsequent works [4], [5], [6]. Initially, the pose transfer task involves the use of source image and keypoints of target pose as input. However, more recent networks have incorporated the parsing map as part of their network input. In fact, some networks [7] don't rely on keypoints at all for similar person image generation tasks; instead, they utilize parsing maps to fuse style features extracted from the source image with the aim of producing a final generated image which preserves the textures of clothes and is visually closer to human perception, thereby enhancing the similarity between the generated image and the real image. Nonetheless, most of these networks predict the parsing map of the generated image based on the target pose. This approach may be weakened if different parts of the body are occluded since it affects their ability to accurately map spatial relationships of poses.

To address the aforementioned challenges, we propose SCRN as a solution. Our approach aims to maintain consistency between the generated image and reference image while gradually transferring the original pose to the target pose using multiple submodules with identical structures that are driven by an attention mechanism. This helps enhance shape consistency between the generated and real person images. Additionally, we incorporate Criss-Cross Attention [8] in each identical module to capture long dependencies, which can provide useful contextual information for visual expression applications. In the first stage of our approach, we input the keypoints of the pose and the conditional semantic distribution into a series of modified Pose Attentional Transfer Blocks (PATBs) [3] to obtain a rough generation result. Afterward, we use pose features and style features to optimize the rough generated results, ultimately achieving the final result. It is important to note that for our method to be effective, it is necessary to calculate losses for both the preliminary results and the final generated results. The main contributions of our paper can be summarized as follows:

- We propose SCRN, which combines two ideas of gradually generating transfer results and transferring style features based on a parsing map. The method can better generate the spatial relationship of the posture while retaining the original image's style characteristics.
- We came up with Coarse Block, a variant of PATB [3]. The conditional semantic distribution and reference semantic distribution generated by prediction are added as input parts at the beginning of the network, so as to restrain the change of posture. Not only that, we also add the Criss-Cross Attention module after each block to enhance the learning ability of the module. Dilated convolution is also introduced to ensure the lightweight of the network.
- In this work, $L1$ loss is performed on the face part alone to produce facial features that better match human perception. The evaluation indicators have improved in all evaluation indicators, which proves the effectiveness of our work.

II. RELATED WORK

A. PERSON IMAGE SYNTHESIS

Person image synthesis refers to the use of computer-generated images to produce realistic depictions of human subjects. This technology has a wide range of applications, such as generating virtual try-on models for clothing and accessories, creating realistic-looking characters for films and games, and even producing digital avatars for social media and virtual reality experiences. Thanks to recent advancements in deep learning techniques, particularly Generative Adversarial Networks (GANs) [9], [10], person image synthesis has made tremendous progress in recent years. GANs allow for the creation of highly realistic images by training a generator network to create images that are indistinguishable from real ones, while simultaneously training a discriminator network to distinguish between real and generated images. With the availability of large datasets and more sophisticated algorithms, person image synthesis is poised to continue advancing at an unprecedented pace.

At the same time, various fashion tasks [11], [12], [13], [14], [15], [16] based on image generation have emerged. Cui et al. propose flexible person generation framework called Dressing in Order (DiOr) [13] which supports 2D

pose transfer, virtual try-on, and several fashion editing tasks. Zhou et al. proposed COutfitGAN [14] to synthesize photo-realistic images of other, complementary, fashion items that are compatible with the given ones. Liu et al. introduced DeepFashion [17], a large-scale multi-task dataset that contains over 800,000 fashion images and covers multiple tasks such as detection, pose estimation, segmentation, and re-identification [18]. This dataset provides researchers with a wide range of application scenarios related to fashion. Ma et al. [19] proposed a pose guided method for person image generation in 2017. This method combines pose estimation and image generation techniques, and generates the corresponding human image by processing the input pose. This method can generate realistic human images in different poses.

B. HUMAN POSE TRANSFER

Human pose transfer, which involves transferring the pose of a person from one image to another, has gained popularity in recent years. The first proposed method for this task was introduced in Ma et al.'s work [19]. Since then, researchers have developed several deep learning-based approaches to simulate human motion, including Liu et al.'s method that uses GAN and convolutional neural networks (CNN) for attitude transfer [20], and Huang et al.'s method that employs style transfer and Adaptive Instance Normalization (AdaIN) technique to achieve high-quality pose transfer in real-time [21]. These methods have been widely used in subsequent works such as Men et al. [4], Zhou et al. [5], and Zhang et al. [6]. The Deformable GANs model proposed by Siarohin et al [22]. employs a novel Deformable Convolutional neural Network (DCN) module to learn how to make subtle shape adjustments to specific regions during generation. This ''deformable G-convolution'' technique allows better handling of objects with complex shapes, such as human body parts, and can make the generated people more natural and realistic. It is worth mentioning that attention mechanism [23], [24] is often used to increase the attention and importance of neural network to input data in pose transfer, so as to effectively realize person image synthesis. Specifically, the attention mechanism allows the neural network to focus only on information in the input data that is relevant to the target task and ignore information that is not relevant to the task. This approach is especially useful when dealing with complex images, which often contain a lot of irrelevant information. In this area, attention mechanisms can be used to identify features that differ between the source domain and the target domain, and thus focus attention on those features. In doing so, the model can better learn the mapping relationship between the source domain and the target domain, and thus more accurately perform pose transfer. Unlike the one-step transfer approach, Huang et al. [3] proposed a pose transfer method that utilizes an attention mechanism. This method involves multiple submodules with identical structures to generate more realistic pose transfer images, through stepwise training, thereby handling complex pose transfer tasks. However, the portrait image generated by this method is based on the keypoints of the pose and may exhibit disconnection between the keypoints in the generated image. This disconnection can occur when there are significant changes in the pose. Also using stacked modules to generate portraits are XingGAN [25] and PoNA [26]. What is different from traditional GAN in that XingGAN introduces a new discriminator structure called ''average pooling discriminator'' to enhance the model's ability to understand global and local information. Different from the existing networks, here, we threaten the pose transfer task into two parts, the spatial representation of the pose and the optimization of the image. Our model can generate a more reasonable pose structure and display a style close to the human senses.

III. METHOD

Our goal is to generate images that are more in line with human perception, based on both the original image and the target pose. To achieve this, we propose SCRN. As shown in Figure 2, the network is divided into two stages: a Coarse Generator and a Refine Generator. The Coarse Generator is responsible for generating rough images I_{crs} that match the original pose P_r. The generated images are then refined using the Refine Generator, which repairs the rough images and makes them more similar to the target image I_t. Additionally, the inputs to the network include the reference image I_r, the reference parsing map S_r, the reference pose P_r, and the target pose P_t. By leveraging these inputs, our model is able to generate high-quality portrait images that closely align with both the original image and the desired pose.

The dataset includes images of the same person in different poses with pose representation. Details are as follows. The pose representation includes 18 human keypoints extracted by Human Pose Estimator (HPE) [27] and provided by [3]. We store the information of 18 joints in text form, and when fed into the network, it generates H $\times$ W $\times$ 18 heat map, where each channel holds the information of a specific joint point of the skeleton. so $P_t \in \mathbb{R}^{H\times W\times 18}$. The parsing map S_r is extracted by Part Grouping Network (PGN) [28] and provided by [6], which saves eight categories (hair, upper clothes, dress, pants, face, upper skin, leg, and background) and processes them into H $\times$ W $\times$ 8 heat map S_r when input to the network. Each category is stored in a separate channel, so $S_r \in \mathbb{R}^{H\times W\times 8}$. In order to generate better facial features, we also use FaceBoxes [29] to extract the bounding box coordinates of the faces in the dataset and save them. We will describe the two stages in detail as follows.

A. COARSE GENERATOR

1) PARSING BLOCK

Accurately generating the pose position is one of the fundamental challenges in human image synthesis. In order

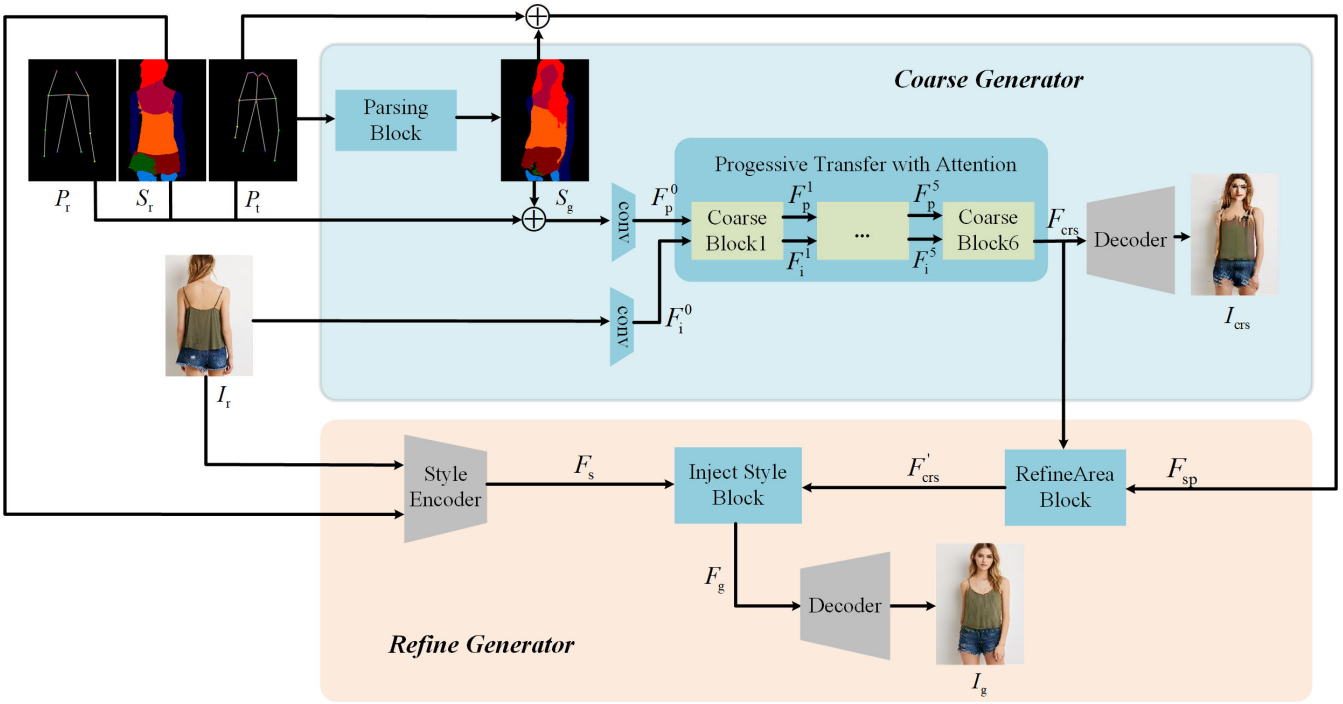


FIGURE 2. Overview of our model.

to tackle this, we utilize a parsing map to constrain pose changes. To generate the parsing map S_g of the target pose, we employ a Parsing Block. First, we embed the source pose P_r, the target pose P_t, and the source parsing S_r into a latent space using an encoder. These inputs are then downsampled through four convolutional layers before residual blocks are applied. This process helps to refine and extract higher-level features from the input data.

2) CRISS-CROSS ATTENTION

The Criss-Cross Attention module is a special attention mechanism for image segmentation tasks in computer vision. This module can help convolutional neural networks better capture the relationship between different spatial locations, thus improving accuracy.

The Criss-Cross Attention module consists of two main steps: First, it extracts the feature vector of the pixel location by using two different convolution kernels at each pixel location. It then converts these two eigenvectors into a two-dimensional matrix to better describe their relationship. We've replaced sigmoid normalization with softmax normalization in this two-dimensional matrix to obtain a matrix of the same size, which indicates the correlation between all pixel locations. Finally, the feature vectors of each pixel position are weighted and averaged using the correlation matrix to generate the final representation.

3) COARSE BLOCK

Based on the work of PATN [3], we introduce a new module called the Coarse Block (CB). Each CB has an identical independent structure. However, when using only pose keypoints as input for the pose feature in the network, the resulting limb disconnects. To overcome this issue, we concatenate S_r, P_r, P_t, and the parsing map S_g generated by the prediction. After applying convolution processing, it becomes the feature F_p^0 which is used as first CB input, meaning F_p^0 contains more information about the pose. Furthermore, we also add Criss-Cross Attention to the module, which better captures the spatial information without bringing a significant increase in computation.

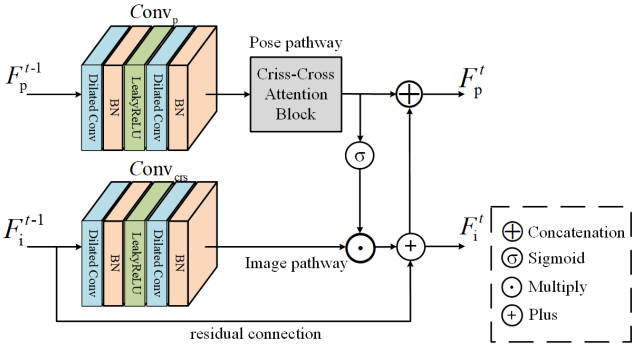


FIGURE 3. Detail of t-th coarse block.

Consider the t-th block, whose inputs are F_p^{t-1} which represents the pose code and F_i^{t-1}, which represents the image code. I_r goes through the convolutional layer and becomes feature F_i^0. As shown in Figure 3, the previous block's output serves as the input for the current block, which is divided into two pathways: the pose pathway and the image pathway. Each of these pathways has a convolutional module with the same

structure but different weights. We denote them as $Conv_p$ and $Conv_{crs}$. Note that we use dilated convolution [30] instead of the standard convolution layer to make the calculation cheaper and have a larger receptive field.

4) POSE PATHWAY AND IMAGE PATHWAY

The basic idea is to hint F_i where to put target patches by pose code F_p. Taking the t-th CB as an example, the output F_p^{t-1} of the (t-1)-th CB is first taken as the input. After convolution processing, the Criss-Cross Attention operation is performed. Finally the sigmoid function is used to obtain the attention matrix M. The pose attention mechanism is introduced to control the details of the generated poses and body parts. This mechanism adjusts the allocation of attention for different stages of generation based on the current pose information, which helps to better align with the desired pose. To summarize, the process can be described as follows:

$$M = \sigma\left(CCA\left(Conv_p\left(F_p^{t-1}\right)\right)\right) \tag{1}$$

where CCA stands for Criss-Cross Attention function. After computing M, we can update F_i^{t-1} and F_p^{t-1} as follows:

$$F_i^t = M \odot Conv_{crs}\left(F_i^{t-1}\right) + F_i^{t-1} \tag{2}$$

where $\odot$ means element-wise multiply. F_p^t is obtained by concatenating F_p^{t-1} and F_i^t in the channel dimension so that the number of channels of F_p^t is the same as the input to the next block. Several CBs (6 in our case) are stacked in this stage. For the output F_i and F_p of the last CB module, F_p was discarded and F_i was retained as the output F_{crs} of the first stage. With Coarse Block, we can better fit the generated pose structure. The F_{crs} will be used as the input of the Refine Generator, while the decoder decodes the F_{crs} to get the preliminary generated image I_{crs}. We will also calculate the loss for I_{crs}.

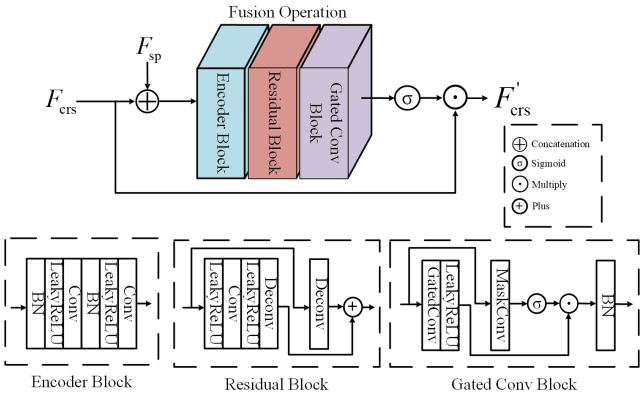


FIGURE 4. Fusion operation of the RefineArea block.

B. REFINE GENERATOR

1) REFINEAREA BLOCK

To make the result more consistent with the ground truth image, we will optimize the output of the Coarse Generator, F_{crs}, with style information. However, and the roughness generated image may differ from the ground truth image in terms of shape, so how to parse I_{crs} with S_r is the main problem to be solved.

RefineArea Block can be regarded as two parts: the first part is encoding P_t and S_g to obtain F_{sp}, and the second part is fusing F_{crs} and F_{sp} to obtain F'_{crs}. The fusion process is shown in Figure 4. Here, gated convolution, a commonly used operation in convolutional neural networks, is introduced, and a gate mechanism regulates the flow of information. In the traditional convolutional neural network, the parameters of each convolution kernel are fixed and cannot be dynamically adjusted according to the input data. However, in gated convolution, each convolution kernel is divided into two parts: a control gate and an activation gate. Control gates are used to decide the importance of input features, while activation gates are used to produce output features. Gated convolution can solve the shape error problem of I_{crs} well. Through the optimization of P_t and S_g, it selectively retains and increases the area that needs to inject style features. F'_{crs} can be expressed as follows:

$$F'_{crs} = \sigma(Fus(F_{crs} \oplus F_{sp})) \odot F_{crs} \tag{3}$$

The Fus here represents the fusion operation in the RefineArea Block and, $\oplus$ represents the concatenation.

2) INJECT STYLE BLOCK

The Inject Style Block plays a crucial role in injecting the texture pattern of the source image into the target pose feature. This is achieved by first extracting the style features F_s from the source image using the style encoder. Incorporating style information can greatly enhance the realism and fidelity of images, as it helps to preserve details and intricacies that would be lost if the texture of the original image was discarded. Without preserving texture, the generated image may appear smooth, blurry, or lacking in detail, failing to capture the essence of the original image.

We use I_r and S_r as the input of the style encoder. The mask generated by S_r enables the encoder to extract the style features of each region accurately. We use five downsampling layers and two residual modules to extract style features. The style information of each region is separately mapped into a 1×256 vector, resulting in F_s. The F'_{crs}, the output by RefineArea Block, is taken as the input of Inject Style Block, and then optimize the F'_{crs} according to the style matched by each channel of each F_s.

C. DISCRIMINATOR

The discriminator is designed to receive two types of input data: real images and fake images generated by the generator. Its primary function is to classify these images and label the real ones as "true" and the generated ones as "false". By doing this, it trains the generator to produce more realistic images by minimizing the difference between the real and fake images, which helps improve its overall accuracy. The discriminator we use consists of nine subsampling layers. I_{crs},

I_g, and I_r are simultaneously fed into the discriminator for each training round.

D. TRAINING

The full loss used to optimize the network consists of the following parts:

$$\mathcal{L} = \lambda_{cor}\mathcal{L}_{cor} + \lambda_{l1}\mathcal{L}_{l1} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{per}\mathcal{L}_{per} + \lambda_{par}\mathcal{L}_{par} + \lambda_{face}\mathcal{L}_{face} \tag{4}$$

where λ_{cor}, λ_{l1}, λ_{adv}, λ_{per}, λ_{par}, and λ_{face} represent the weights of the loss values of each part, and these parameters are used to balance the loss values of each part, so that the network converges faster.

Same as the previous method, we introduce the correspondence loss which is used to constrain the generated feature F_g with target image. The formula is as follows:

$$\mathcal{L}_{cor} = \left\| F_g - \phi_i\left(I_t\right) \right\|_2 \tag{5}$$

where ϕ_i represents the pre-trained VGG19 [31] model, and we use part of the layers of VGG19 to calculate the loss. The i here represents the number of layers used.

In order to make I_g more like I_t, we need to measure their similarity at the pixel level. So we use $L1$ losses. It is important to note that not only do we calculate losses between I_g and I_t, but we also calculate losses between I_{crs} and I_t. In this way, we can make the Coarse Generator generate I_g better. The $L1$ loss in this paper is calculated as follows:

$$\mathcal{L}_{l1} = \left\| I_g - I_t \right\|_1 + \left\| I_{crs} - I_t \right\|_1 \tag{6}$$

The adversarial loss is calculated by the discriminator D. It punishes the distribution differences between the generated (fake) image I_g and the rough image I_{crs}, respectively, and the expected (real) target image.

$$\mathcal{L}_{adv} = \mathbb{E}\left[\log\left(1 - D\left(I_g\right)\right)\right] + \mathbb{E}\left[\log\left(1 - D\left(I_{crs}\right)\right)\right] + \mathbb{E}\left[\log D\left(I_t\right)\right] \tag{7}$$

The function for perceptual loss [32] gauges the variance between feature representations of the target and generated images at intermediate layers, promoting their similarity. This approach results in visually appealing outcomes that uphold the original image's global structure and semantic content, rather than simply matching pixel values. It can be expressed by the formula:

$$\mathcal{L}_{per} = \left\| \Theta\left(I_g\right) - \Theta\left(I_t\right) \right\|_1 \tag{8}$$

where Θ denotes the pre-trained AlexNet [33].

Parsing map loss is calculated by $L1$ loss, which can be written as:

$$\mathcal{L}_{par} = \left\| S_g - S_t \right\|_1 \tag{9}$$

During training, we find it harder to generate reasonable images of faces, so we compute the loss specifically for faces. The bounding box coordinates of the face are generated and saved by FaceBoxes [29]. The loss is calculated by cropping the faces of I_g and I_t according to bounding boxes. The formula is as follows:

$$\mathcal{L}_{face} = \left\| C\left(I_g\right) - C\left(I_t\right) \right\|_1 \tag{10}$$

where C represents the function to crop the face.

IV. EXPERIMENT

A. EXPERIMENT SETUP

1) DATASET

Our model is trained using the In-shop Clothes Retrieval Benchmark, part of the DeepFashion [17] dataset. This particular dataset comprises 52,712 high-resolution images featuring fashion models. In order to ensure effective training and testing, images depicting the same individual in identical attire have been paired together. We use the dataset splits provided by [3]. There are 101, 966 pairs in the training set and 8, 570 pairs in the testing set.

2) METRICS

To assess the performance of our model, we rely on various evaluation metrics such as Structure Similarity Index Measure (SSIM), Peak Signal to Noise Ratio (PSNR), Fréchet Inception Distance (FID), and Learned Perceptual Image Patch Similarity (LPIPS). These metrics are used to measure the difference between the generated image and the target image. SSIM is a widely used evaluation metric to compare the structural similarity of two images. It considers not only the brightness and contrast of the image, but also the structural information, which was firstly used in [19]. However, SSIM can be insensitive to certain image transformations. Another important metric we use is PSNR, where a higher value indicates a smaller difference between the generated and real images. FID is another useful metric that measures the difference between the generated and real images by calculating the distance between their feature representations in the feature space.Finally, we use LPIPS, a neural network-based metric that learns feature representations to evaluate image similarity from the human perspective. LPIPS is more closely aligned with human perception and can provide a more accurate reflection of the differences between images compared to traditional evaluation metrics.

3) IMPLEMENTATION DETAILS

Our method is implemented in PyTorch, and the graphics card used is an Nvidia 3090. We set the batch size to 6 and took 120k iterations to get the final result. The Adam optimizer [34] has $\beta 1 = 0.9$ and $\beta 2 = 0.999$. The learning rate is initially set to 0.001 and decays to 0 after 120k iterations. The weights for each part of the loss values are: $\lambda_{cor} = 300$, $\lambda_{l1} = 5$, $\lambda_{adv} = 2$, $\lambda_{per} = 3$, $\lambda_{par} = 1$ and $\lambda_{face} = 1$.

B. COMPARRISONS WITH STATE-OF-THE-ART MODELS

1) QUANTITATIVE COMPARISON

We compare our model with several state-of-the-art models, including PATN [3], ADGAN [4], PINet [35], PoNA [26],

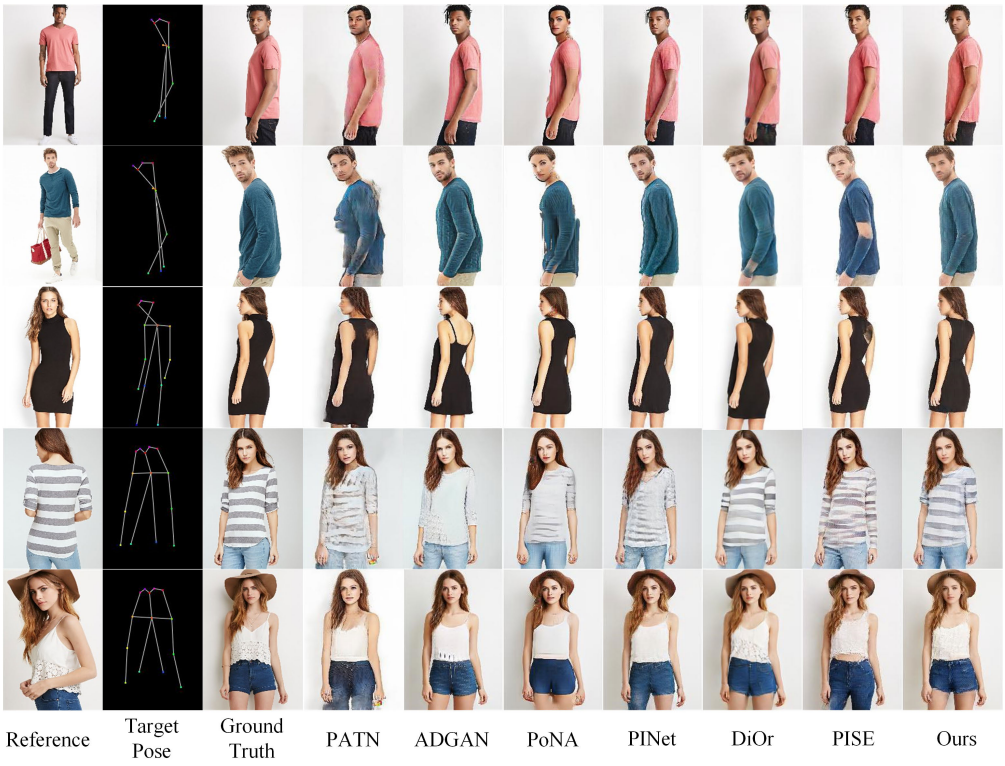


FIGURE 5. Qualitative comparisons with state-of-the-art methods. From left to right are the results of PATN, ADGAN, PINet, PoNA, DiOr, PISE and ours, respectively.

FIGURE 6. Face comparisons with state-of-the-art methods. From left to right are the results of PATN, ADGAN, PINet, PoNA, DiOr, PISE and ours, respectively.

DiOr [13] and PISE [6]. The generated images used to test the evaluation metrics are provided by the corresponding authors, and the width and height of the images are uniformly 256×176. Therefore, the generated image of some models with width and height of 256×256 will be cropped. The networks we used for comparison use the same dataset split,

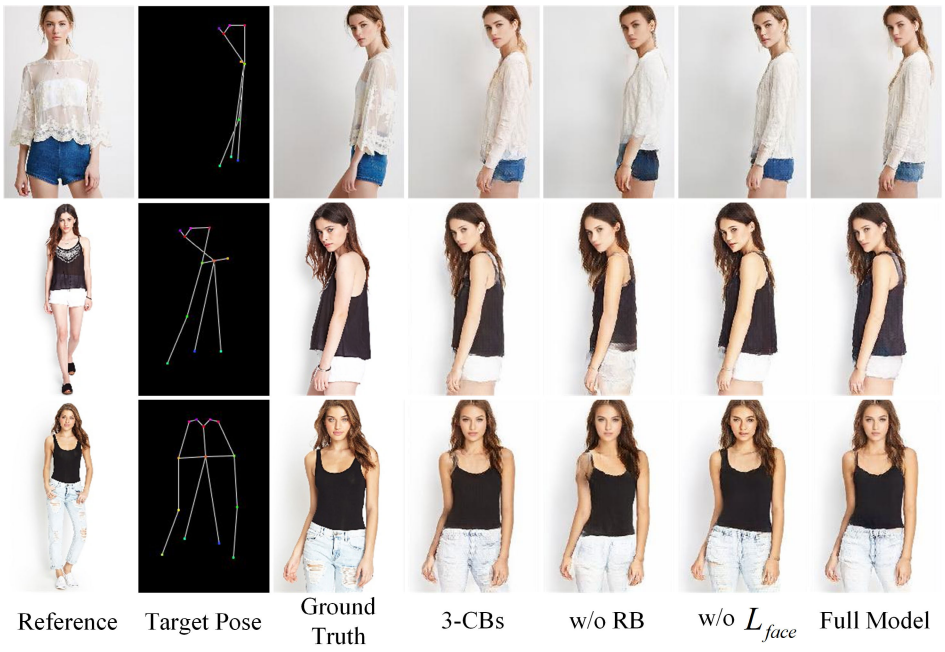


FIGURE 7. Qualitative comparisons in ablation study.

TABLE 1. Quantitative comparisons with state-of-the-art methods.

Model	SSIM↑	FID↓	PSNR↑	LPIPS↓
PATN	0.7069	17.1699	15.5684	0.2861
ADGAN	0.7459	12.2656	16.7704	0.2255
PINet	0.7415	11.4476	16.9132	0.2167
PoNA	0.7487	14.7148	16.7831	0.2504
DiOr	0.7534	11.6465	17.2553	0.2206
PISE	0.7417	**11.0102**	16.9081	0.2078
Ours	**0.7564**	11.0900	**17.3098**	**0.1873**

TABLE 2. Quantitative comparisons in ablation study.

Model	SSIM↑	FID↓	PSNR↑	LPIPS↓
3-CBs	0.7481	11.4279	17.1161	0.1923
w/o RB	0.7326	12.9100	16.3635	0.2202
w/o L_{face}	0.7521	11.1462	17.2145	**0.1837**
Full Model	**0.7564**	**11.0900**	**17.3098**	0.1873

so these evaluation metrics are measured on the full test set. Since PATN [3] does not provide results for running on the test set, we directly use the source code and trained model released by the author of PATN [3] to test metrics. The quantitative comparison result is shown in Table 1. Among them, DiOr [13] has shown excellent performance, but in LPIPS its performance is poor. It can be seen from the table that our model has a slight improvement in SSIM, indicating that it has made a more accurate prediction for the details, texture, structure and other aspects of the image. Our model obtains the best results on PSNR and LPIPS, second in FID, proving that the generated images more aline with human visual perception.

2) QUALITATIVE COMPARISON

We provide the generated results of comparison with state-of-the-art network in Figure 5, The images used for display are all released by their respective authors. It is obvious that the ability of PATN [3] and ADGAN [4] to preserve the texture of the source image is weak. In the example of the last row in Figure 5, PATN [3] and ADGAN [4] do not capture the information about the hat, so the generated portrait does not wear a hat. We believe that the attention mechanism used in the PATN [3] model has some limitations, because the attention mechanism only focuses on a part of the area in the image, so it may not capture all the information in the image in some cases. Some images generated by PISE [6] do not generate the boundary of clothes and skin accurately, and we guess that this is because the "Per-rigion Normalization" module used by PISE [6] may mistakenly confuse the clothing texture with the skin texture in some cases.

However, the result shows that our model generates images with accurate structure while preserving texture features well. It is worth mentioning that when generating full-body images, the face area is small compared to the image, and the face of other models will become blurred or deformed. Due to the introduction of L_{face}, our model can still generate reasonable facial features even in this case. See Figure 6 for a specific example. By the way, looking through all the images

generated with the test set, it can be seen that the quality of the images we generate is more stable.

C. ABLATION STUDY

In this section, we trained several ablation networks, which are variants of our network, to verify the effectiveness of our improvements.

3-Coarse Blocks(3-CBs). Coarse Blocks are used to change a person's posture progressively. Its structure is simple and it relies on attention mechanisms to drive it. This article used six Coarse Blocks. In this section, we will also test the model with 3 Coarse Blocks.

Without RefineArea Block(w/o RB). A RefineArea Block is used to selectively retain and add style features to areas where they need to be injected. The target pose and generated parsing map are encoded in this module. The crudely generated image is fusion manipulated with the encoded result.

Without L_{face}(w/o L_{face}). In the full model, the face part of the ground truth image and the face part of the generated image was clipped with the same corresponding bounding box frame and the loss L_{face} was calculated so that a more reasonable face could be generated by introducing L_{face}. In this section, The model does not adopt the L_{face} loss defined in Equation (10).

Full Model. We trained our model with all components we proposed.

We trained all our ablation models using the same setting. The quantitative comparison and qualitative comparison can be seen in Table 2 and Figure 7, respectively. It can be found that the model with all components achieved the best performance overall on the evaluation metrics.

The model utilizes only 3 CBs that result in poor LPIPS performance, indicating that the model lacks efficient space transformation ability due to CB reduction. Each CB module restricts attitude change within a small range to avoid losing features and prevent excessive image alteration in one step. It is worth noting that theoretically, better results can be achieved by increasing the number of CBs. The original PATN [3] uses nine similar modules, and six CBs are used in this model considering the amount of computation.

The model without RefineArea Block is significantly lower than the full model in all metrics. The poor SSIM and LPIPS indicate that the structure between the generated image and the original image is not similar enough, and the poor FID indicates that the generated result is not realistic enough. According to the qualitative comparison result, we can also find that since there is no RefineArea Block, the shape of the resulting image I_g will largely depend on I_{crs}. Since we inject the style feature at element-wise multiply, it is necessary to add the shape information with RefineArea Block to not restrict the I_g shape to the I_{crs}.

FIGURE 8. Failure cases caused by rare poses, rare clothes, and male to female ratio.

Although the model without L_{face} performs similarly to the full model in terms of evaluation metrics and even outperforms the full model in LPIPS, the face generated by the full model is superior to that of the model without L_{face}. We speculate that the small size of the face area compared to the entire image contributes to the marginal difference in the evaluation metrics. Hence, it does not have a significant impact on calculating the evaluation metrics.

D. FAILURE CASES

Although our model achieves good results in most cases, there are still some limitations. As shown in Figure 8. We classify limitations into the following three categories:

Rare poses. These poses are rare even in the training set. For example, it is difficult to find the sitting position in the entire dataset, so when generating the sitting position, the generated result is extremely biased towards the composition of the training set, resulting in the wrong postures.

Rare clothes. Some clothes are not very common in life, even in the DeepFashion dataset. This will result in clothing that is very different from the source image.

Generated males. There were 101, 966 pairs of images in the training set, including 90, 152 pairs of female images and 11, 814 pairs of male images. The male-to-female ratio in the dataset is nearly 9:1, a severe data imbalance that sometimes results in male faces resembling women.

In fact, all of the above three limitations can be attributed to insufficient datasets or data imbalance. We believe that expanding the data set can effectively alleviate these problems.

V. CONCLUSION

In this paper, we present a novel approach for human pose transfer that involves the generation of rough images followed by refinement while preserving the spatial structure and style features of the human body. Our approach consists of two stages: first, a sketch is generated by incrementally changing the pose through several modules with identical structures. In the second stage, the previously generated image is fused with keypoints and semantic distribution information, and style features are injected to generate the final image. Additionally, we incorporate a facial loss calculation to produce more realistic facial features. Our experiments demonstrate that our approach produces images that are more consistent with human perception and closely resemble real images. Furthermore, our ablation study confirms the effectiveness of our approach. Later we will try to use this method for subject transfer which consists of transferring not only the pose but also the appearance and background.

REFERENCES

[1] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "VITON: An image-based virtual try-on network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7543–7552.

[2] J. Walker, K. Marino, A. Gupta, and M. Hebert, "The pose knows: Video forecasting by generating pose futures," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3352–3361.

[3] Z. Zhu, T. Huang, B. Shi, M. Yu, B. Wang, and X. Bai, "Progressive pose attention transfer for person image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2342–2351.

[4] Y. Men, Y. Mao, Y. Jiang, W.-Y. Ma, and Z. Lian, "Controllable person image synthesis with attribute-decomposed GAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5083–5092.

[5] X. Zhou, M. Yin, X. Chen, L. Sun, C. Gao, and Q. Li, "Cross attention based style distribution for controllable person image synthesis," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*. Tel Aviv, Israel: Springer, Oct. 2022, pp. 161–178.

[6] J. Zhang, K. Li, Y.-K. Lai, and J. Yang, "PISE: Person image synthesis and editing with decoupled GAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7978–7986.

[7] N. Dufour, D. Picard, and V. Kalogeiton, "Scam! Transferring humans between images with semantic cross attention modulation," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*. Tel Aviv, Israel: Springer, Oct. 2022, pp. 713–729.

[8] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.

[9] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[10] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.

[11] N. Rostamzadeh, S. Hosseini, T. Boquet, W. Stokowiec, Y. Zhang, C. Jauvin, and C. Pal, "Fashion-Gen: The generative fashion dataset and challenge," 2018, *arXiv:1806.08317*.

[12] Y. Lang, Y. He, J. Dong, F. Yang, and H. Xue, "Design-GAN: Cross-category fashion translation driven by landmark attention," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 1968–1972.

[13] A. Cui, D. McKee, and S. Lazebnik, "Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14618–14627.

[14] D. Zhou, H. Zhang, Q. Li, J. Ma, and X. Xu, "COutfitGAN: Learning to synthesize compatible outfits supervised by silhouette masks and fashion styles," *IEEE Trans. Multimedia*, early access, Jun. 23, 2022, doi: 10.1109/TMM.2022.3185894.

[15] D. Zhou, H. Zhang, K. Yang, L. Liu, H. Yan, X. Xu, Z. Zhang, and S. Yan, "Learning to synthesize compatible fashion items using semantic alignment and collocation classification: An outfit generation framework," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Sep. 15, 2022, doi: 10.1109/TNNLS.2022.3202842.

[16] I. Choi, S. Park, and J. Park, "Generating and modifying high resolution fashion model image using StyleGAN," in *Proc. 13th Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Oct. 2022, pp. 1536–1538.

[17] 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.

[18] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline in vitro," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3774–3782.

[19] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," in *Proc. 31st Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[20] W. Liu, Z. Piao, J. Min, W. Luo, L. Ma, and S. Gao, "Liquid warping GAN: A unified framework for human motion imitation, appearance transfer and novel view synthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct./Nov. 2019, pp. 5903–5912.

[21] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1510–1519.

[22] A. Siarohin, E. Sangineto, S. Lathuilière, and N. Sebe, "Deformable GANs for pose-based human image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3408–3416.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Dec. 2017, pp. 6000–6010.

[24] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7354–7363.

[25] H. Tang, S. Bai, L. Zhang, P. H. Torr, and N. Sebe, "XingGAN for person image generation," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*. Glasgow, U.K.: Springer, Aug. 2020, pp. 717–734.

[26] K. Li, J. Zhang, Y. Liu, Y.-K. Lai, and Q. Dai, "PoNA: Pose-guided non-local attention for human pose transfer," *IEEE Trans. Image Process.*, vol. 29, pp. 9584–9599, 2020.

[27] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1302–1310.

[28] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, and L. Lin, "Instance-level human parsing via part grouping network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 770–785.

[29] S. Zhang, X. Wang, Z. Lei, and S. Z. Li, "FaceBoxes: A CPU real-time and accurate unconstrained face detector," *Neurocomputing*, vol. 364, pp. 297–309, Oct. 2019.

[30] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[32] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*. Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 694–711.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[34] J. L. Long, N. Zhang, and T. Darrell, "Do convnets learn correspondence?" in *Proc. 27th Adv. Neural Inf. Process. Syst.*, vol. 27, Dec. 2014, pp. 1601–1609.

[35] J. Zhang, X. Liu, and K. Li, "Human pose transfer by adaptive hierarchical deformation," *Comput. Graph. Forum*, vol. 39, no. 7, pp. 325–337, 2020.

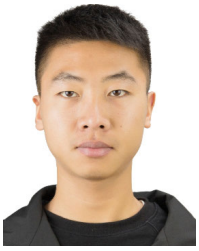
HAN MO was born in Guilin, Guangxi, China, in 1999. He received the bachelor's degree in engineering from Guizhou University, in 2022, where he is currently pursuing the master's degree in information and communication engineering. During the bachelor's degree, he won the second-class scholarship and the outstanding three good student. During the master's degree, he won the first-class scholarship for freshmen. His research interests include computer vision, image generatio, and internet technologies.

YOUJU PENG was born in Bijie, Guizhou, China, in 1998. She received the bachelor's degree in engineering from Guizhou University, in 2021, where she is currently pursuing the master's degree in information and communication engineering.

During the bachelor's degree, she won the first-class scholarship, the national inspirational scholarship, the outstanding three good student, and provincial outstanding graduate. During the master's degree, she won the second-class scholarship and third-class scholarship for freshmen. Her research interests include computer vision, AI, wild mushrooms classification, and internet technologies.

YANG XU received the bachelor's degree from Guizhou University, in 2003, and the Ph.D. degree in engineering from the Institute of Modern Materials and Mechanics, Chinese Academy of Sciences, in 2008.

He is currently a Master Tutor with Guizhou University. He is also an associate professor. His current research interests include big data acquisition, the Internet of Things technology, and embedded technology.



GUIDONG XU was born in Bijie, Guizhou, China, in 1995. He received the bachelor's degree in engineering from the Nanjing Institute of Technology, in 2019. He is currently pursuing the master's degree in communication engineering with Guizhou University. He won a second-class scholarship while studying for the master's degree. His research interests include computer vision, machine learning, and internet technology.