REMOT: A Region-to-Whole Framework for Realistic Human Motion Transfer

Quanwei Yang†
yangquanwei@mail.ustc.edu.cn
University of Science and Technology of China

Xinchen Liu*
Wu Liu
liuxinchen1@jd.com
liuwu1@jd.com
JD Explore Academy

Hongtao Xie
htxie@ustc.edu.cn
University of Science and Technology of China

Xiaoyan Gu
guxiaoyan@iie.ac.cn
Institute of Artificial Intelligence, Chinese Academy of Sciences

Lingyun Yu
yuly@mail.ustc.edu.cn
University of Science and Technology of China

Yongdong Zhang
zhdy73@ustc.edu.cn
University of Science and Technology of China

ABSTRACT
Human Video Motion Transfer (HVMT) aims to, given an image of a source person, generate his/her video that imitates the motion of the driving person. Existing methods for HVMT mainly exploit Generative Adversarial Networks (GANs) to perform the warping operation based on the flow estimated from the source person image and each driving video frame. However, these methods always generate obvious artifacts due to the dramatic differences in poses, scales, and shifts between the source person and the driving person. To overcome these challenges, this paper presents a novel REMOT framework based on GANs. To generate realistic motions, the REMOT adopts a progressive generation paradigm: it first generates each body part in the driving pose without flow-based warping, then composites all parts into a complete person of the driving motion. Moreover, to preserve the natural global appearance, we design a Global Alignment Module to align the scale and position of the source person with those of the driving person based on their layouts. Furthermore, we propose a Texture Alignment Module to keep each part of the person aligned according to the similarity of the texture. Finally, through extensive quantitative and qualitative experiments, our REMOT achieves state-of-the-art results on two public benchmarks.

KEYWORDS
Human Motion Transfer; Video Generation; Generative Adversarial Network

1 INTRODUCTION
The task of Human video Motion Transfer (HVMT) is to synthesize a realistic video where a source person performs desired driving motion [32, 49]. Specifically, given the source image and a driving video, it needs to make the person in the source image imitate the action of the person in the driving video. In recent years, this task has received intense research because of its wide range of applications, such as digital man [37] and person re-identification [35, 37]. However, the quality of generated motion videos is still far from satisfactory and needs further improvement.

With the development of deep learning, remarkable achievements have been made in many fields [2, 6, 9, 10, 22, 23, 25, 29, 30, 36, 41, 43, 55]. Among them, the rise of Generative Adversarial Networks (GANs) [6] makes synthesized images or videos become more and more realistic [18, 45, 46, 52, 53]. Existing HVMT methods also adopt the GAN-based framework to achieve realistic results [4]. According to the generalization of the model, current HVMT methods can be divided into personalized HVMT and general-purpose HVMT. Personalized methods [4, 47, 50] implicitly embed the global appearance of a specific person into the generator, then synthesize the images of this person conditioned on the driving poses. Although such methods can synthesize realistic video frames, these methods require a large amount of video data of a specific person for training. Moreover, this type of method has no generalization

†This work is done when Quanwei Yang is an intern at JD Explore Academy.
*Xinchen Liu is the corresponding author.
ability for different persons, which has to train an exclusive model for each subject.

To overcome the huge data requirement of individual subjects and improve the generalization of personalized methods, recent works mainly focus on general-purpose methods [5, 12, 15, 24, 38, 44, 49]. This type of method aims to learn a model that can be adapted for the generation of unseen persons. Even if the generated images may not be satisfactory, the results can be improved by simply fine-tuning the existing models instead of re-training them. Therefore, this paper is also concentrated on the general-purpose human motion transfer which has a wider range of applications.

Although general-purpose methods have many advantages, they still face several challenges. First of all, because different parts of one person, such as the face, the body, the pants, etc., usually have different textures, it is difficult to model these complex textures with a uniform GAN. For example, some direct transfer methods [44] use a GAN to generate the whole person by fusing the appearance and pose features, which hardly preserve local details well, as shown in Figure 1 (a). Second, the drastic variations of poses and viewpoints between the source person and the driving person bring great challenges to human motion transfer. Some recent methods adopt warping flows, such as the optical flow [49], the transform flow [24], etc., to achieve pose transfer, as shown in Figure 1 (b). Nevertheless, the generated frames still have significant artifacts caused by the errors of flow estimation. Last but not the least, the misalignment of scales and positions between the source person and the driving person can also degrade the quality of generated videos.

To overcome the above challenges, this paper presents a novel REgion-to-whole framework for realistic human MOTion Transfer, named REMOT. Figure 1 (c) demonstrates the conceptual architecture of the proposed progressive framework. The REMOT first predicts the layout (semantic parsing masks) of the human body conditioned on the driving pose. Then a region generator is designed to generate each body part individually. Moreover, we propose a Global Alignment Module (GAM) to keep the scales and positions of the source person and the driving person aligned for accurate feature fusion. Finally, a GAN-based whole compositor is utilized to integrate the parts and refine the coarse synthesized person image. To preserve the local details of the source person, we design a Texture Alignment Module (TAM) to align the features of the source person image and the generated frames. In summary, our key contributions are as follows:

- We propose a region-to-whole framework, named REMOT, for the general-purpose HVMT task. REMOT discards the flow-based warping operation and takes a progressive generation paradigm, which can better deal with dramatic variations of poses by gradually generating person frames from parts to the whole.
- To overcome the differences in scales and positions between the source person and the driving ones, we propose a Global Alignment Module to adapt the scale and position of the source person to those of the driving person.
- A Texture Alignment Module is proposed to align the features of the source image and the initially generated image to preserve more details like textures of clothes and edges of the bodies.

We conduct extensive experiments on the iPER Dataset [24] and SoloDance Dataset [49]. Experimental results show that our model achieves the state-of-the-art both quantitatively and qualitatively.

## 2 RELATED WORK

### 2.1 Image-to-Image Translation

Our method is related to the image-to-image translation task that converts one source image to a target image conditional on a specific style, resolution, contents, etc [14]. Existing image-to-image translation methods usually adopt the Conditional Generative Adversarial Networks [21, 33] to take the specific image (e.g., semantic segmentation map, sketch, edge map, etc.) as input for image generation. Such models can generate the desired image by editing the properties of the input image. Therefore, Compared with unconstrained GANs, image-to-image translation has a wider range of applications. For example, Pix2pix [14] used a conditional discriminator to make the generated image and the input image match as much as possible, so it can generate realistic corresponding images. Based on pix2pix, pix2pixHD [45] generated high-resolution images by using multi-scale generators and discriminators. Vid2vid [46] was aimed at the temporal inconsistency of the generated videos. By adding optical flow constrained to the generator and discriminator, vid2vid can generate temporally coherent videos.

The personalized motion transfer is to translate the pose map to the corresponding RGB image, where the person’s appearance is
implicitly embedded into the generator. Therefore, many image-to-image translation works can be used for personalized motion transfer. However, such methods cannot be directly applied to HVMT for the generation of varied poses and precise details of persons.

2.2 Human Video Motion Transfer

Personalized HVMT. On the basis of the image-to-image translation methods, some works have specific improvements based on the characteristics of this task. In order to conform to the size of the source person, EDN [4] matched the poses between the driving person and the source person by designing global pose normalization. In addition, EDN designed a face GAN to generate a more precise face region. DIW [47] encoded non-uniformly sampled pose maps of the past ten frames to capture appearance change details and refine pose features of the single frame. As mentioned earlier, personalized methods suffer from the inability to generalize to unseen person generations and need massive computing resources.

General-purpose HVMT. To overcome the limitations of personalized HVMT, recent studies focus on general-purpose HVMT. In order to better preserve appearance details, some works adopt the warp operation. These methods focusing on the animation of general objects, such as FOMM [38] and MRAA [39], utilized predicted keypoints or regions to compute the flow field between the source image and the driving image, and then warped the feature of the source image to obtain the target image. However, these methods do not take temporal consistency into consideration, which results in the temporal inconsistency of the synthesized video. LW-GAN [24] warped the source information in both image and feature spaces based on the SMPL model. However, the SMPL model [31] was only suitable for smooth human bodies, it cannot represent the human body with complex clothes. Different from warping the feature of the source image, CZF [49] estimated the optical flow of the clothing regions and directly warped the clothing region according to optical flow. Unfortunately, when the driving pose is greatly different from the source pose, such methods usually failed due to inaccurate flow estimations. Different from the above methods, our proposed method does not use warp operation. This design can avoid complex transformation calculations and can better cope with arbitrary poses.

In addition, some methods consider the diversity of different regions of the human body and treat each region of the human body separately. SSF [5] encoded each region of the body separately, then integrated features of different regions to directly generate the final target image. SHUP [1] first transformed each part of the source person to the spatial position of the driving pose, then directly generated the whole person image. Moreover, PGTM [12] predicted the UV map [8] of the source person conditioned on driving pose, then warped the feature of the source image to obtain the target image. However, predicting dense UVs in different poses accurately is challenging. StylePeople [7] used parameterized human body model SMPL-X [34] to model the pose and shape of subjects, then used neural rendering to render region appearance details. Our method also encodes each region of the source image separately. However, unlike these methods, we directly generate corresponding target region images, then use the whole composition network to integrate the generated regions for generating the final target image. Compared with directly generating the whole person image, generating the region image first can better preserve the appearance details.

3 THE PROPOSED METHOD

Given the source image \( I_s \) and a driving video \( V_d = \{ I'_d^t \}, t = 1, 2, \cdots, N \), where \( I'_d^t \) represents the driving frame at time \( t \) of the driving video, our goal is to generate a realistic video in which the person of the source image imitates the same motion as the driving video. We can formulate this task as

\[
\hat{I}'_d^t = \mathcal{F} \left(I'_d^{t-1}, I'_d^t \right), \quad t = 1, 2, \cdots, N, \tag{1}
\]

where \( \mathcal{F}(\cdot, \cdot) \) refers to generative model, \( N \) represents the number of video frames, \( \hat{I}'_d^t \) represents the generated image.

Figure 2 is an overview of our pipeline. Our proposed method consists of two networks: the Region Generation Network and the Whole Composition Network. For the Region Generation Network, given the pose of the driving image \( P_{d, t} \) and the human body layout of the source image \( L_s \), we first generate the layout of the source person conditioned on the driving pose \( \hat{L}_d^t \) through Layout GAN. Then the Region GAN takes the generated mask of each region \( \hat{L}_d^t \) and the corresponding region of the source person after global alignment \( L_s^i \), as input, and obtains each source person region image \( \hat{R}^t_{d, i} \) in the driving pose. Finally, in the Whole Composition Network, we refine the coarse person image \( \hat{F}_G^t_{d, i} \), by aligning the features of the source person with the Texture Alignment Module. In addition, we add the background image to the generated person image to obtain the final target image \( \hat{I}'_d^t \). In the following, we will introduce the details of these two networks.

3.1 Region Generation Network

As mentioned before, the reason why we first generate each source person region in the driving pose is that different regions of the human body have significantly different texture patterns. It’s difficult to directly generate the whole human body. To achieve this, we need to first generate the layout of the source person in the driving pose through Layout GAN, as shown in Figure 2 (a).

Layout GAN. In order to represent the human pose, we first use the human keypoint detection method [3, 27, 28] to predict the 2-D keypoint \( K \in \mathbb{R}^{2 \times 25} \) of the human body. Then according to the predefined connection strategy, we can get the pose connection map \( P_{d, t}^i \in \mathbb{R}^{2 \times H \times W} \) of the driving person, where \( H \times W \) is the resolution of the image. In order to represent the human body layout, we use the human parsing methods [20, 26, 54] to obtain the 18-channel semantic segmentation maps of the person image. Further, according to the similarity of textures in various regions, we merge the semantic segmentation maps into six classes, i.e., head, top, bottom, shoes, limbs, and background. So we can get the body layout \( L_s \in \mathbb{R}^{6 \times H \times W} \) of the source person image and \( L_{d, t}^i \) of the driving person image.

For the network structure, we also use vid2vid [46] as in [49]. Specifically, we use three encoders \( E_1^t, E_2^t \) and \( E_3^t \) to encode the source body layout \( L_s \), the concatenated driving poses connection map \( \{ P_{d, t-2}^i, P_{d, t-1}^i, P_{d, t}^i \} \) and the concatenated results previously generated \( \{ \hat{L}_{d, t-2}^i, \hat{L}_{d, t-1}^i \} \) to obtain the feature maps \( F_1^t, F_2^t \) and \( F_3^t \).
WCN takes the output of the RGN where + weight $\omega$ video frames coherent temporally, not for the driving image and the generated results from previous moment and make generated transformation of nated along the channel dimension. Warp multiplication, respectively. Then decoder $D^i_1$ can be formulated as $F = \sum_{i=1}^{n} E^i_1(\hat{L}_d^i, P_d^i) + \sum_{i=1}^{n} E^i_2((\hat{L}_d^i, \hat{L}_d^{i-1}))$. (2c)

where $+$, $*$ represent element-wise addition and element-wise multiplication, respectively. $\{\}$ indicates that the inputs are concatenated along the channel dimension. Warp($I, o$) refers to the affine transformation of $I$ according to the optical flow $O$. It should be emphasized that the warp operation in the formula is only to integrate the generated results from previous moment and make generated video frames coherent temporally, not for the driving image and the source image.

Region GAN. After obtaining the human body layout $\hat{L}_d^i$ of the source person conditioned on driving pose from Layout GAN, Region GAN is proposed to generate five person regions conditioned on the driving pose.

For the network structure, we still adopt the vid2vid structure. Since Region GAN is only for generating raw region images of the source person, here we only use one generator to generate five regions of the human body. This can not only save computational resources but also prevent the model from over-fitting. Specifically, we use the encoder $E^i_1$ to encode the i-th region mask $\hat{L}_d^i \in \mathbb{R}^{1 \times 2W \times W}$ of the human body to obtain the feature map $F_d^i$. In addition, we use another encoder $E^i_2$ to encode the i-th region of the source person image and the previous generations to obtain the feature map $F_{d,i}$. Here we directly concatenate $\hat{L}_d^{i-1}, \hat{L}_d^{i-2}$ and $L_{s,i}$ along the channel dimension to utilize the previous generations for enhancing the temporal consistency of appearance details. The original Region GAN can be expressed as $\hat{L}_d^i = D^i_l \left( E^i_1(\hat{L}_d^i, \hat{L}_{d,i}), E^i_2(\hat{L}_d^{i-1}, \hat{L}_{d,i}, L_{s,i}) \right)$.

Figure 2: The pipeline of our proposed REMOT. It consists of two networks: (a) Region Generation Network (RGN) and (b) Whole Composition Network (WCN). The RGN takes the source image and driving frame as input to obtain each region of the source person in the driving pose. (The blue arrows represent the data preprocessing: pose detection, layout detection.) The WCN takes the output of the RGN $\hat{F}_d$, the source foreground image after global alignment $\hat{F}_A$, and concatenated human layout with pose map $\{\hat{L}_d^i, P_d^i\}$ as input, then generates the final target frame.
Global Alignment Module. The scale and spatial position of the person in the video may vary greatly, so the scales and positions of \( I_{d,i}^{t} \) and \( I_{s} \) may not match, then the extracted features \( F_{d,i}^{t} \) and \( F_{s} \) are also not aligned. Therefore, simply adding \( F_{d,i}^{t} \) and \( F_{s} \) does not fuse the features well. To scale it, we propose the Global Alignment Module (GAM), a simple but effective Affine transformation to transform the \( I_{d} \) before feeding it to \( E_{d} \), so that the scale and location of \( I_{d} \) are consistent with the \( I_{s}^{t} \).

See the green rectangle of Figure 2 (a). First, we obtain the human masks \( M_{d,i}^{t} \) and \( M_{s} \) according to \( I_{d,i}^{t} \) and \( L_{s} \), respectively. According to the human mask, we can obtain the foreground image \( F_{G}s \) of the source image. We express the process of aligning the foreground of the source foreground image with the following formulation:

\[
\overline{F_{G}s} = \begin{bmatrix} R & 0 \\ 0 & R \end{bmatrix} F_{G}s + \begin{bmatrix} c_{x} \\ c_{y} \end{bmatrix},
\]

where \( \overline{F_{G}s} \) represents the globally aligned foreground image of the source image. \( R \) represents the scaling factor, which is obtained according to \( H_{d}, H_{s}, H_{d} \) and \( H_{s} \) represent the height of \( M_{d} \) and \( M_{s} \), respectively. \( c = [c_{x}, c_{y}]^{T} \) represents the position offset, which can be obtained by \( c_{x} = -c_{y} \), where \( c_{x} \) and \( c_{y} \) represent the center position coordinates of \( M_{d}^{t} \) and \( M_{s} \), respectively. So the Region GAN can be expressed as:

\[
\hat{I}_{d,i}^{t} = D_{r}^{t} \left( E_{r}^{t}(I_{d,i}^{t}) + E_{r}^{t}\left(I_{d,i}^{t-1}, I_{d,i}^{t-2}, \text{GAM}(I_{s,i})\right)\right).
\]

The loss of Region GAN is the sum of the loss of five synthesized region images, and each loss consists of three items, e.g., the reconstruction loss, the perceptual loss [16], and the adversarial loss.

Reconstruction loss. It directly constrains the generated image \( \hat{I}_{d,i}^{t} \) in the pixel space to be closer to ground truth \( I_{d,i}^{t} \). Compared with the L2 loss, L1 loss can focus on the subtle differences of the image. Its formulation is as follows:

\[
L_{rec,i}^{t} = \left\| \hat{I}_{d,i}^{t} - I_{d,i}^{t} \right\|_{1}.
\]

Perceptual loss. It regularizes the generated image \( \hat{I}_{d,i}^{t} \) and ground truth \( I_{d,i}^{t} \) to be closer in multi-dimensional feature space. The perceptual loss includes the feature content loss and feature style loss, which can be expressed as

\[
L_{per,i}^{t} = \sum_{j=1}^{n} \left( \| \phi_{j} \left( \hat{I}_{d,i}^{t} \right) - \phi_{j} \left( I_{d,i}^{t} \right) \|_{1} + \| G_{\phi}^{j} \left( \hat{I}_{d,i}^{t} \right) - G_{\phi}^{j} \left( I_{d,i}^{t} \right) \|_{1} \right),
\]

where \( j \) represents the \( j \)-th layer of the pre-trained VGG-19 [40] model, and \( G \) represents the Gram matrix of features.

Adversarial loss. It aims to force the synthesized images to conform to a similar distribution of real images. In order to make the network focus on multi-scale image details, we also use the multi-scale conditional discriminator proposed in pix2pixHD [14]. It takes a synthesized image and the corresponding human region mask as input. Its expression is

\[
L_{Gado,i}^{t} = \sum D_{k}(\hat{I}_{d,i}^{t}, I_{d,i}^{t}).
\]

Therefore, the entire loss function of the generator is shown below.

\[
L_{G}^{t} = \sum_{i=1}^{5} \left( \lambda_{rec}L_{rec,i}^{t} + \lambda_{per}L_{per,i}^{t} + L_{Gado,i}^{t} \right),
\]

where \( \lambda_{rec} \) and \( \lambda_{per} \) are the weights for the reconstruction loss and perceptual loss, respectively.

For the discriminator, its loss function is

\[
L_{D} = \sum \mathbb{E}[\log D_{k}(I_{d,i}^{t}, I_{d,i}^{t})] + \mathbb{E}[\log (1 - D_{k}(\hat{I}_{d,i}^{t}, I_{d,i}^{t}))].
\]

3.2 Whole Composition Network

After generating each region image, we can obtain the foreground image of the target image using the following formula:

\[
F_{G}^{t} = \sum_{i=1}^{5} \hat{I}_{d,i}^{t}.
\]

However, this is only the sum of the individual regions of the human body generated by Region GAN, each region is still insufficient in connections and some details. In addition, we still need to add the background to the generated person image. Therefore, the Whole Generation Network is proposed to further enhance the details of the person, improve the unreasonable areas and add the background image to the person image.

As shown in Figure 2 (b), for the Whole Composition Network, we use three encoders to extract information. a) We use encoder \( E_{3}^{w} \) to encode the concatenated \( \{I_{d,i}^{t-2}, I_{d,i}^{t-1}\} \) to obtain feature map \( F_{3}^{w} \in \mathbb{R}^{c \times h \times w} \). Three adjacent frames as input can implicitly improving the temporal consistency of generated frames. b) We use the encoder \( E_{2}^{w} \) to encode concatenated \( \{F_{G}^{t-2}, F_{G}^{t-1}, F_{G}^{t, d,r}\} \) to get the feature map \( F_{2}^{w} \in \mathbb{R}^{c \times h \times w} \). c) In order to exploit the true foreground information of the source person image, we use the encoder \( E_{1}^{w} \) to encode \( F_{G}s \) to get the feature map \( F_{1}^{w} \in \mathbb{R}^{c \times h \times w} \). Because \( I_{d,i}^{t}, I_{d}^{t}, F_{G}^{t, d,r} \) are all in the driving pose, the obtained feature maps \( F_{1}^{w}, F_{2}^{w}, F_{3}^{w} \) can be directly added for feature fusion.

Texture Alignment Module. However, \( F_{G}^{t, d,i} \) and \( F_{G}s \) are the same person with the same textures but having different poses. So the Texture Alignment Module (TAM) is proposed to better fuse feature maps \( F_{1}^{w} \) and \( F_{2}^{w} \) for refining local details. TAM calculates the texture similarity between each position of \( F_{1}^{w} \) and each position of \( F_{2}^{w} \), then adds the aligned \( F_{1}^{w} \) to \( F_{2}^{w} \).

As shown in the blue rectangle of Figure 2 (b), we first reshape \( F_{1}^{w} \) into \( H_{1} \) in \( \mathbb{R}^{c \times h \times w} \), and we can get \( H_{2} \) in \( \mathbb{R}^{c \times h \times w} \) by \( F_{2}^{w} \) in the same way. Then we calculate the similarity of each position in \( H_{1} \) to each position in \( H_{2} \) to get the affinity matrix \( A \in \mathbb{R}^{h \times w} \). Here we use the cosine distance to represent the texture similarity. The calculation formula is as follows:

\[
A = \text{softmax} \left( \frac{H_{1,i,j}^{T} \cdot H_{2,i,j}}{||H_{1,i,j}|| ||H_{2,i,j}||} \right), \quad i, j = 1, 2, \ldots, h, w,
\]

where \( \cdot \) means matrix multiplication, \( H_{1,i,j} \) represents the feature of \( H_{1} \) at the \( i \)-th position, and \( H_{2,i,j} \) represents the feature of \( H_{2} \) at the \( j \)-th position. Next, we multiply \( H_{1} \) by the affinity matrix to get the alignment vector, then add it point-to-point with \( H_{2} \), finally reshape it to the original dimension to get the fused feature map.
Table 1: Quantitative comparison to state-of-the-art methods on the iPER dataset. ↑ indicates the higher is better, correspondingly, ↓ indicates the lower is better.

| Methods   | SSIM ↑ | PSNR ↑ | LPIPS ↓ | FID ↓ | TCM ↑ |
|-----------|--------|--------|---------|-------|-------|
| EDN [4]   | 0.840  | 23.39  | 0.076   | 56.29 | 0.361 |
| FSV2V [44]| 0.780  | 20.44  | 0.110   | 110.99| 0.184 |
| LWGAN [24]| 0.825  | 21.43  | 0.091   | 77.99 | 0.197 |
| C2F [49]  | 0.849  | 24.27  | 0.072   | 55.07 | 0.687 |
| Ours      | 0.856  | 25.33  | 0.065   | 53.04 | 0.793 |

\[ F_{i,t}^w \in \mathbb{R}^{C \times h \times w}, \text{which can be formulated as} \]
\[
F_{i,t}^w = \text{reshape} \left( H_1 \cdot A + H_2 \right). \tag{13}
\]

After getting the fused texture feature \( F_{i,t}^w \), we add it to the pose feature \( F_{i,t}^w \) and send it to decoder \( D_{i,t}^w \) to get the final foreground image \( \hat{F}_G \). The entire foreground image generation process can be expressed as

\[ \hat{F}_G = D_{i,t}^w \left( F_{i,t}^w + \text{TAM}(F_{i,t}^w, F_{i,t}^w) \right). \tag{14} \]

Although a realistic person image is synthesized, we still need to add a reasonable background to the foreground person image. Therefore, we send the additive feature map \( F_{i,t}^w \) and \( F_{i,t}^w \) into decoder \( D_{i,t}^w \) to get the soft-mask \( m \), then multiply it with \( BG \) to get the corresponding background image. Finally, we add the background to the foreground image to get the final image. Its calculation formula can be expressed as

\[ \hat{I}_d = m \times BG + \hat{F}_G. \tag{15} \]

It should be noted that we also use the same strategy as LayoutGAN, using the predicted optical flow of adjacent frames to further refine the generated image. In addition to using the same loss function as Region GAN, we also design another discriminator specifically for the facial region to generate realistic faces.

3.3 Training and Inference

Training. In the training phase, in order to get the paired training data as supervision information, we select a forward frame in a video as the source image and this video as the driving video. Before training our model, we first use pre-trained models to extract the poses and the human parsing maps of frames. We train our REMOT step by step. First, we train Layout GAN and Region GAN separately for 10 epochs. Then we train the Whole Composition Network for 10 epochs using the output of Region GAN.

Inference. In the inference stage, the choice of driving video is not limited, as long as it is a clear video of any solo person. For convenience, our model can perform end-to-end inference.

4 EXPERIMENTS

This section consists of four subsections. We first introduce the Implementation Details and Datasets (Sec. 4.1). Then compare our method with some representative existing methods quantitatively and qualitatively (Sec. 4.2 and Sec. 4.3). Finally, we show the effectiveness of each module through the ablation study (Sec. 4.4). Each subsection is described in detail below.

4.1 Implementation Details and Datasets

Implementation Details. Like [49], all frames are resized and cropped to 256x192 before training the models. We use Opencv [3] to get the 2-D keypoints of the human body, use SCHP [20] to get the human parsing, and use the optical flow obtained by FlowNet2 [13] as ground truth. We set \( \lambda_{src} \), \( \lambda_{per} \) in Eq(10) to 10 for balancing several losses. We use Adam optimizers [19] with learning rate of 0.0002, where \( \beta_1 = 0.5, \beta_2 = 0.999 \) on Nvidia Tesla P40.

Dataset. In order to comprehensively evaluate the performance of our proposed model, we choose two public datasets for training and testing.

iPER Dataset. The iPER dataset was proposed by [24], which was collected in the laboratory environment. The dataset includes 30 actors with a total of 206 videos. Its action types of an actor include A-pose and random actions. The scale of iPER is relatively large and the actions are relatively simple. Same as the original protocol of iPER, 164 videos are used for training and the remaining 42 videos for testing.

SoloDance Dataset. Different from the iPER dataset, SoloDance dataset contains 179 solo dance videos in real scenes collected online [49], and each video has 300 frames. The dataset has 143 subjects, and its action types are mainly dances such as hip-hop and modern dance. Compared with the iPER dataset, the actions in this dataset are more complex and the types of clothes are more diverse. However, this dataset is not as large as the iPER dataset. Following the setting of [49], we also use 153 videos for training and 26 videos for testing.

Compared Approaches. We select several representative state-of-the-art HVMT methods for comparison with our proposed approach. They are personalized methods END [4], directly generation method FSV2V [44], feature warping method LWGAN [24], and image warping method C2F [49]. We also use 3,000 video frames for each subject to train the exclusive EDN model, and the training strategies of other methods are the same as our method.

4.2 Quantitative Results

Metrics. In order to comprehensively compare the quality of generated images, we use four image-level evaluation metrics: SSIM [48], PSNR, Learned Perceptual Image Patch Similarity (LPIPS) [56], Fréchet Inception Distance (FID) [11] and one metric for evaluating the temporal consistency of generated videos, i.e., Temporal Consistency Metric (TCM) [51]. SSIM measures the structural similarity between the generated image and the real image in pixel space. PSNR evaluates the generation quality based on pixel-wise
errors. LPIPS assesses the perceptual distance based on feature vectors. FID represents the Inception [42] feature distance between two image sets. As for TCM, it evaluates the temporal consistency of generated videos by calculating the distance between the current video frame and the warped neighborhood frame with optical flow.

A quantitative comparison of our method with other methods on the iPER dataset is shown in Table 1. It shows that our proposed method outperforms other methods in terms of image quality and temporal consistency. EDN trains the exclusive model for each specific person, so it achieves satisfactory results on all metrics. Because FSV2V directly generates the whole person image, resulting in unreasonable person images, it performs the worst among all methods. For LWGAN, there is no constraint on temporal consistency, resulting in jitters in the generated human video. Therefore, LWAGN does not perform well on the TCM metric. Since half of the videos of rotating actions in the iPER dataset, it is difficult for C2F to handle such a situation with drastic pose differences. Thanks to proposed region-to-whole strategy, our method can be applied to more situations, thus achieving the best results on all metrics.

Table 2 provides the comparison of our method with other methods on the SoloDance dataset. It can be seen that we have achieved the best results on the Mask-SSIM, Mask-PSNR, LPIPS and TCM. (Mask means masking the background of the image according to the human parsing map) This shows that our method has advantages in the quality of generated person images and the temporal consistency of generated videos. The reason why we don’t achieve the best results on other metrics may be that the scale of the SoloDance dataset is not large enough for training the model composed of entirely GANs compared to the iPER dataset. In addition, C2F directly retains the clothes of the source image based on predicted optical flow, which makes it achieve the best results on the FID metric. However, this also leads to generating unreasonable person images in the case of dramatic pose differences between the source person and the driving person. (It can be seen in the areas marked by the yellow dotted boxes of Figures 3 and 4.) Overall, our method can achieve the state-of-the-art on both two datasets.

### 4.3 Qualitative Results

As shown in Figures 3 and 4, we randomly select three video frames of different poses from synthesized videos on iPER and SoloDance datasets. Although EDN can generate realistic images in some cases, unseen poses often lead to unreasonable person images (Marked with green boxes in Figures 3 and 4). Therefore, EDN is limited by the pose diversity of the particular person in the training set. It can be seen that due to the inaccuracy of the SMPL predicted by HMR [17], LWGAN will generate uneven texture and jitter (Marked with red boxes in Figures 3 and 4). For C2F, when the poses of the driving image and the source image are different dramatically, optical flow prediction becomes inaccurate. So the synthesized clothes cannot be warped precisely, which results in unrealistic person images. (Marked with yellow boxes in Figures 3 and 4) This may seriously affect human visual perception. Compared with other methods, our method can better handle the situation of drastic pose changes while preserving the details. In addition, the person images generated by our model generally have clearer faces. This is due to our progressive model in which the initial face image provides an important template for the final clear face.

### 4.4 Ablation Study

In order to verify the roles of the Global Alignment Module (GAM) and Texture Alignment Module (TAM) on the generated results and the effectiveness of the Whole Composition Network (WCN), we perform the ablation experiments on the iPER and SoloDance datasets. The variants of the framework are as follows:

- **RGN w/o GAM.** It means that the foreground image is generated without using the GAM.
- **REMOT w/o WCN.** It means that the foreground image is generated through the RGN using the GAM.
- **REMOT w/o TAM.** It means that the REMOT model directly adds the feature $F_1^w$ to the feature $F_2^w$ point-to-point, instead of using the TAM.
- **REMOT.** It refers to our proposed complete REMOT model.

The experimental results are shown in Tables 3 and 4. Since the RGN only generates person foreground images, we only calculate the Mask-SSIM and Mask-PSNR. Experimental results in Table 3 show that the GAM does not significantly improve the quality of the generated images on the iPER dataset. This is because in iPER videos, the motion of actors is relatively simple and the scale and position of actors remain unchanged, which means that the source image and the driving image are originally in a weak alignment. However, the GAM plays a greater role because of the large range of actions of the subjects in the SoloDance dataset. Obviously, on both datasets, WCN can significantly improve the quality of generated images. Compared with direct feature addition for fusion, TAM can better fuse the features of the generated raw foreground image with the features of the source image.

As can be seen from Figure 5, WCN plays an important role in enhancing facial details and clothing textures. Compared with the direct addition of features, TAM is more beneficial to generate realistic person images.

### 5 CONCLUSION

In this paper, we propose a progressive human motion transfer framework, namely REMOT, which gradually generates person images for the source video. The proposed framework, namely REMOT, which gradually generates person images for the source video. The framework consists of a Global Alignment Module (GAM) and a Texture Alignment Module (TAM). The GAM globally aligns the source and driving images, while the TAM aligns the features of the generated images point-to-point. The proposed framework outperforms other methods in terms of image quality and temporal consistency.

| **Table 3**: Results of ablation experiments of our method on the iPER dataset. * indicates Mask-SSIM and Mask-PSNR. |
|---|---|---|---|---|---|
| **Method** | **SSIM**↑ | **PSNR**↑ | **LPIPS**↓ | **FID**↓ | **TCM**↑ |
| RGN w/o GAM | 0.892 | 20.94 | - | - | - |
| REMOT w/o WCN | 0.895 | 21.32 | - | - | - |
| REMOT w/o TAM | 0.928 | 26.62 | 0.074 | 55.23 | 0.715 |
| REMOT | 0.952 | 28.87 | 0.065 | 53.04 | 0.793 |

| **Table 4**: Results of ablation experiments of our method on the SoloDance dataset. |
|---|---|---|---|---|---|
| **Method** | **SSIM**↑ | **PSNR**↑ | **LPIPS**↓ | **FID**↓ | **TCM**↑ |
| RGN w/o GAM | 0.921 | 24.56 | - | - | - |
| REMOT w/o WCN | 0.937 | 25.54 | - | - | - |
| REMOT w/o TAM | 0.941 | 26.25 | 0.053 | 64.37 | 0.723 |
| REMOT | 0.953 | 27.89 | 0.045 | 53.29 | 0.788 |
Figure 3: Qualitative results on the iPER dataset. Visible artifacts are marked with colored dotted boxes.

Figure 4: Qualitative results on the SoloDance dataset. Visible artifacts are marked with colored dotted boxes.

Figure 5: Results of ablation experiments.

Images from region to whole. REMOT abandons the warping operation, which avoids inaccurate flow estimation due to drastic variations of poses. Compared with directly generating the whole person image, the region-to-whole strategy makes the HVMT task easier and generated person images more realistic. Moreover, we propose a Global Alignment Module to match the size and position of the source person with those of the driving person. Furthermore, we propose the Texture Alignment Module to align the features of the source image with the generated image to preserve more details. Experiments on the iPER and SoloDance datasets show that our proposed approach achieves state-of-the-art results. In future work, we will try 3D information of the human body for HVMT, which can provide powerful prior knowledge such as 3D shapes and poses of the human body.

ACKNOWLEDGMENTS
This work is supported by the National Key R&D Program of China under Grant No. 2020AAA0103800, the National Nature Science Foundation of China (62121002, 62022076, U1936210, 62102127), the Fundamental Research Funds for the Central Universities under Grant WK3480000011, the Youth Innovation Promotion Association Chinese Academy of Sciences (Y2021122), and the Hefei Postdoctoral Research Activities Foundation (BSH202101).
REFERENCES

[1] Guha Balakrishnan, Amy Zhao, Adrian V. Dalca, Frédéric Durand, and John V. Guttag. 2018. Synthesizing Images of Humans in Unseen Poses. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 8340–8348.

[2] Xiaochun Cao, Hao Zhang, Xiaojie Guo, Si Liu, and Dan Meng. 2015. SLED: Semantic Label Embedding Dictionary Representation for Multilabel Image Annotation. IEEE Trans. Image Process. 24, 9 (2015), 2746–2759.

[3] Tinne O. Cso, Gino H. Hafilo, Tomasz Simon, Shih-En Wei, and Yaser Sheikh. 2021. OpenPose: Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. IEEE Trans. Pattern Anal. Mach. Intell. 43, 1 (2021), 172–186.

[4] Caroline Chan, Shary Ginose, Tinghui Zhuo, and Alexei A. Efros. 2019. Everybody Dance Now. In IEEE/CVF International Conference on Computer Vision (ICCV). 5932–5941.

[5] Oran Gafni, Oron Ashual, and Lior Wolf. 2021. Single-Shot Freestyle Dance Time Style Transfer and Super-Resolution. In ACM International Conference on Multimedia (MM). 11066–11076.

[6] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 3431–3440.

[7] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Szegedy. 2015. Rethinking the Inception Architecture for Computer Vision. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2818–2826.

[8] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Sze. 2017. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 770–778.

[9] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Bryan Catanzaro, and Yangqing Jia. 2016. Residual Networks: Deep Learning Without Overfitting. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 694–711.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In Advances in Neural Information Processing Systems (NIPS). 5998–6008.

[11] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In International Conference on Learning Representations (ICLR).

[12] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Sze. 2017. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2818–2826.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In Advances in Neural Information Processing Systems (NIPS). 5998–6008.

[14] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In International Conference on Learning Representations (ICLR).

[15] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. High-Resolution Image Synthesis and Semantic Manipulation With Conditional GANs. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 8798–8807.

[16] Tuanfeng Y. Wang, Duygu Ceylan, Krishna Kumar Singh, and Niboy J. Mitra. 2021. Dance in the Wild: Human Motion Inference With Neural Dynamic Appearance Synthesis. In International Conference on 3D Vision (3DV). 628–277.

[17] Zhiquan Liu, Chen Lian, Xiaokui Duan, Yuxing Gao, and Tao Mei. 2019. Social Relation Recognition From Videos via Multi-Scale Spatial-Temporal Reasoning. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 3564–3572.

[18] Zhiquan Liu, Chen Lian, Xiaokui Duan, Yuxing Gao, and Tao Mei. 2019. Social Relation Recognition From Videos via Multi-Scale Spatial-Temporal Reasoning. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 3564–3572.

[19] Zhiquan Liu, Chen Lian, Xiaokui Duan, Yuxing Gao, and Tao Mei. 2019. Social Relation Recognition From Videos via Multi-Scale Spatial-Temporal Reasoning. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 3564–3572.

[20] Zhiquan Liu, Chen Lian, Xiaokui Duan, Yuxing Gao, and Tao Mei. 2019. Social Relation Recognition From Videos via Multi-Scale Spatial-Temporal Reasoning. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 3564–3572.

[21] Zhiquan Liu, Chen Lian, Xiaokui Duan, Yuxing Gao, and Tao Mei. 2019. Social Relation Recognition From Videos via Multi-Scale Spatial-Temporal Reasoning. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 3564–3572.

[22] Zhiquan Liu, Chen Lian, Xiaokui Duan, Yuxing Gao, and Tao Mei. 2019. Social Relation Recognition From Videos via Multi-Scale Spatial-Temporal Reasoning. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 3564–3572.

[23] Zhiquan Liu, Chen Lian, Xiaokui Duan, Yuxing Gao, and Tao Mei. 2019. Social Relation Recognition From Videos via Multi-Scale Spatial-Temporal Reasoning. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 3564–3572.

[24] Zhiquan Liu, Chen Lian, Xiaokui Duan, Yuxing Gao, and Tao Mei. 2019. Social Relation Recognition From Videos via Multi-Scale Spatial-Temporal Reasoning. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 3564–3572.

[25] Zhiquan Liu, Chen Lian, Xiaokui Duan, Yuxing Gao, and Tao Mei. 2019. Social Relation Recognition From Videos via Multi-Scale Spatial-Temporal Reasoning. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 3564–3572.

[26] Zhiquan Liu, Chen Lian, Xiaokui Duan, Yuxing Gao, and Tao Mei. 2019. Social Relation Recognition From Videos via Multi-Scale Spatial-Temporal Reasoning. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 3564–3572.

[27] Zhiquan Liu, Chen Lian, Xiaokui Duan, Yuxing Gao, and Tao Mei. 2019. Social Relation Recognition From Videos via Multi-Scale Spatial-Temporal Reasoning. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 3564–3572.

[28] Zhiquan Liu, Chen Lian, Xiaokui Duan, Yuxing Gao, and Tao Mei. 2019. Social Relation Recognition From Videos via Multi-Scale Spatial-Temporal Reasoning. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 3564–3572.

[29] Zhiquan Liu, Chen Lian, Xiaokui Duan, Yuxing Gao, and Tao Mei. 2019. Social Relation Recognition From Videos via Multi-Scale Spatial-Temporal Reasoning. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 3564–3572.

[30] Zhiquan Liu, Chen Lian, Xiaokui Duan, Yuxing Gao, and Tao Mei. 2019. Social Relation Recognition From Videos via Multi-Scale Spatial-Temporal Reasoning. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 3564–3572.
[49] Dongxu Wei, Xiaowei Xu, Haibin Shen, and Kejie Huang. 2021. C2F-FWN: Coarse-to-Fine Flow Warping Network for Spatial-Temporal Consistent Motion Transfer. In AAAI Conference on Artificial Intelligence (AAAI). 2852–2860.

[50] Zhuoqian Yang, Wentao Zhu, Wayne Wu, Chen Qian, Qiang Zhou, Bolei Zhou, and Chen Change Loy. 2020. TransMoMo: Invariance-Driven Unsupervised Video Motion Retargeting. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 5305–5314.

[51] Chun-Han Yao, Chia-Yang Chang, and Shao-Yi Chien. 2017. Occlusion-aware Video Temporal Consistency. In ACM International Conference on Multimedia (MM). 777–785.

[52] Lingyun Yu, Hongtao Xie, and Yongdong Zhang. 2022. Multimodal Learning for Temporally Coherent Talking Face Generation With Articulator Synergy. IEEE Trans. Multim. 24 (2022), 2950–2962.

[53] Lingyun Yu, Jun Yu, Mengyan Li, and Qiang Ling. 2021. Multimodal Inputs Driven Talking Face Generation With Spatial-Temporal Dependency. IEEE Trans. Circuits Syst. Video Technol. 31, 1 (2021), 203–216.

[54] Dan Zeng, Yuhang Huang, Qian Bao, Junjie Zhang, Chi Su, and Wu Liu. 2021. Neural Architecture Search for Joint Human Parsing and Pose Estimation. In IEEE/CVF International Conference on Computer Vision (ICCV). 11365–11374.

[55] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. 2021. PyMAF: 3D Human Pose and Shape Regression with Pyramidal Mesh Alignment Feedback Loop. In IEEE/CVF International Conference on Computer Vision (ICCV). 11426–11436.

[56] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 586–595.

[57] Jinkai Zheng, Xincheng Liu, Wu Liu, Lingxiao He, Chenggang Yan, and Tao Mei. 2022. Gait Recognition in the Wild With Dense 3D Representations and a Benchmark. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 20228–20237.