HUMAN MOTIONFORMER: TRANSFERRING HUMAN MOTIONS WITH VISION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Human motion transfer aims to transfer motions from a target dynamic person to a source static one for motion synthesis. An accurate matching between the source person and the target motion in both large and subtle motion changes is vital for improving the transferred motion quality. In this paper, we propose Human MotionFormer, a hierarchical ViT framework that leverages global and local perceptions to capture large and subtle motion matching, respectively. It consists of two ViT encoders to extract input features (i.e., a target motion image and a source human image) and a ViT decoder with several cascaded blocks for feature matching and motion transfer. In each block, we set the target motion feature as Query and the source person as Key and Value, calculating the cross-attention maps to conduct a global feature matching. Further, we introduce a convolutional layer to improve the local perception after the global cross-attention computations. This matching process is implemented in both warping and generation branches to guide the motion transfer. During training, we propose a mutual learning loss to enable the co-supervision between warping and generation branches for better motion representations. Experiments show that our Human MotionFormer sets the new state-of-the-art performance both qualitatively and quantitatively.

1 INTRODUCTION

Human Motion Transfer, which transfers the motion from a target person's video to a source person, has grown rapidly in recent years due to its substantial entertaining applications for novel content generation Wang et al. (2019); Chan et al. (2019). For example, a dancing target automatically animates multiple static source people for efficient short video editing. Professional actions can be transferred to celebrities to produce educational, charitable, or advertising videos for a wide range of broadcasting. Bringing static people alive suits short video creation and receives growing attention on social media platforms.

During motion transfer, we expect the source person to redo the same action as the target person. To achieve this purpose, we need to establish an accurate matching between the target pose and the source person (i.e., each body part skeleton in a pose image matches its corresponding body part in a source image), and use this matching to drive the source person with target pose (i.e., if the hand skeleton in the target pose is raised, the hand in the source person should also be raised). According to the degree of difference between the target pose and the source person pose, this matching can be divided into two types: **global** and **local**. When the degree is large, there is a large motion change between the target pose and the source person, and the target pose shall match a distant region in the source image (e.g., the arm skeleton of the target pose is distant from the source man arm region in Fig. 1(b)). When the degree is small, there are only subtle motion changes, and the target pose shall match its local region in the source image (e.g., the arm skeleton of the target pose is close to the source woman arm region in Fig. 1(b)). As the human body moves non-rigidly, large and subtle motion changes usually appear simultaneously. The local and global matching shall be conducted simultaneously to ensure high-quality human motion transfer.

Existing studies Ren et al. (2020; 2021); Tao et al. (2022); Zhao & Zhang (2022) leverage 2D human body keypoints Cao et al. (2017); Xiu et al. (2018) as the initial pose representations and introduce image-to-image translation Isola et al. (2017) or predictable warping fields (e.g., optical flow, affine transformations) for geometric matching. Since the 2D human body keypoints are sparse and the

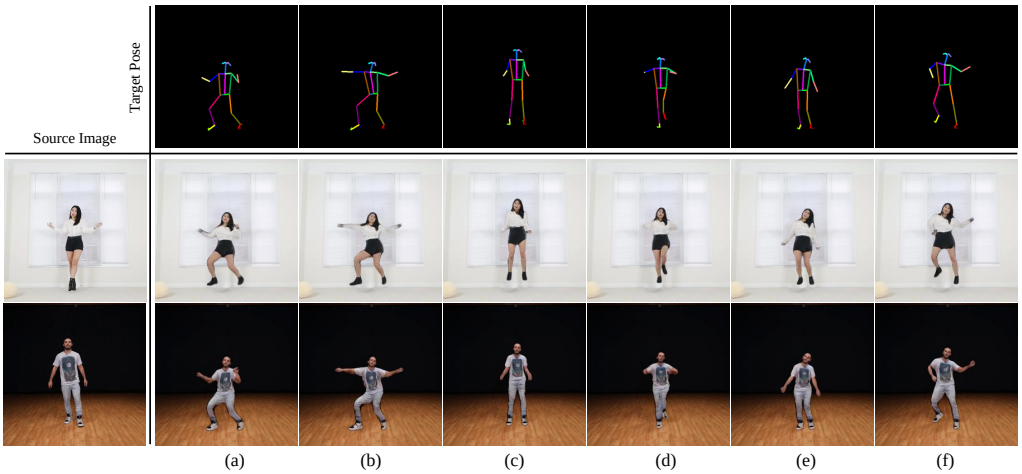


Figure 1: Human motion transfer results. Target pose images are in the first row, and two source person images are in the first column. Our MotionFormer effectively synthesizes motion transferred results whether the poses in the above two images differ significantly or not.

geometric matching is captured by CNNs, it is hard to capture global matching as the CNN receptive field is limited. As a result, artifacts occur and details are missing in the transferred results when facing large motion changes between the target pose and source person. The alternative 3D methods Neverova et al. (2018); Gafni et al. (2021); Huang et al. (2021a) introduce DensePose Alp Güler et al. (2018) or parametric body mesh Loper et al. (2015) as human representations to perform pixel-wise matching. They can globally match target pose and source images by aligning two humans into one 3D model. However, the off-the-shelf Densepose and 3D models suffer from background interference and partial occlusion, limiting the image alignment quality for dense pixel-wise matching.

Is there a proper way to simultaneously utilize robust 2D human body keypoints and build global and local matching? Fortunately, recent methods Dosovitskiy et al. (2020); Liu et al. (2021; 2022); Dong et al. (2022) demonstrate that the Vision Transformers (ViTs) can capture global dependencies in visual recognition. Inspired by this design, we combine the advantages of CNNs and Transformers and propose a Vision-Transformers-based framework named MotionFormer to model the visual correspondence between the source person image and the target pose image. As shown in Figure 2, our MotionFormer consists of two encoders and one decoder. The encoders extract feature pyramids from the source person image and the target pose image, respectively. These feature pyramids are sent to the decoder for accurate matching and motion transfer. In the decoder, there are several decoder blocks and one fusion block. Each block consists of two parallel branches (i.e., the warping and generation branches). In the warping branch, we predict a flow field to warp features from the source person image, which preserves the information of the source image to achieve high-fidelity motion transfer. Meanwhile, the generation branch produces novel content that cannot be directly borrowed from the source appearance to improve photorealism further. Afterward, we use a fusion block to convert the feature output of these two branches into the final transferred image.

The feature matching result dominates the flow field and generation content. Specifically, we implement feature matching from two input images via convolutions and cross-attentions in each branch. The tokens from the source image features are mapped into Key and Value, and the tokens from the target pose image features are mapped into Query. We compute the cross-attention map based on the Key and Query. This map reflects the global correlations between two input images. Then, we send the output of cross-attention process into a convolution to capture locally matched results. Thanks to the accurate global and local feature matching, our MotionFormer can improve the performance of both the warping and generation processes. Moreover, we propose a mutual constrain loss to enable these two branches to supervise each other during training, which helps them to benefit from each other and facilitates the final fusion for high-quality motion transfer. In the test phase, our method generates a motion transfer video on the fly given a single source image without training a person-specific model or fine-tuning. Some results generated by our method can be found in Fig. 1, and we show more video results in the supplementary files.

2 RELATED WORKS

Human Motion Transfer. Most human motion transfer works are built upon an image-to-image Isola et al. (2017); Wang et al. (2018b) or video-to-video Wang et al. (2018a; 2019) translation framework. With this framework, some methods Chan et al. (2019); Esser et al. (2018); Dong et al. (2018); Han et al. (2019a); Ma et al. (2017); Ren et al. (2021; 2020) set the off-the-shelf 2D body keypoints Cao et al. (2017); Xiu et al. (2018) as a condition to animate the source person image. And some methods Siarohin et al. (2019a;b; 2021) use the unsupervised 2D body keypoints to extract motion representations. Therefore, these methods generalize well to a wider range of objects(i.e., animals). The 2D body keypoints can represent the body motion correctly, but this representation is sparse and the CNN models used by these methods have limited receptive fields, which leads to inaccurate global visual correspondence between this motion representation and the source image. To overcome this limitation, recent approachesNeverova et al. (2018); Grigorev et al. (2019); Huang et al. (2021a); Shysheya et al. (2019); Liu et al. (2019); Ma et al. (2018) project the source person and the target person into a unified 3D space (i.e., DensePose Alp Güler et al. (2018) and SMPL Loper et al. (2015)) to capture pixel-level correspondences, which help render the appearance from source to target in global perspective. However, compared to 2D keypoint representations, DensePose and SMPL models are less accurate and may produce large misalignments in complex scenes. In this paper, we design the first Vision-Transformer-based generation framework for human motion transfer. With merely 2D keypoints guiding the target motion, we capture large motion deformations with globally attended appearance and motion features and yield state-of-the-art performance. Unlike many previous methods that trained on a single video Chan et al. (2019); Shysheya et al. (2019) or need fine-tuning Huang et al. (2021a); Liu et al. (2019); Zakharov et al. (2019) to achieve higher perceptual quality, our method works in a one-shot fashion that directly generalizes to unseen identities.

Vision Transformer. Transformer Vaswani et al. (2017) has become increasingly popular in solving the computer vision problems, such as object detection Carion et al. (2020); Liu et al. (2021); Touvron et al. (2021), segmentation Wang et al. (2021a); Zheng et al. (2021); Cao et al. (2021b), inpainting Peng et al. (2021); Wan et al. (2021), image generation Cao et al. (2021a); Lee et al. (2021); Jiang et al. (2021); Esser et al. (2021), restoration Chen et al. (2021b); Wang et al. (2021b); Liang et al. (2021), image classification Wang et al. (2021a); Huang et al. (2021b); Ramachandran et al. (2019); Liu et al. (2021); Wu et al. (2021); Dong et al. (2022); Wu et al. (2020); Srinivas et al. (2021); Chen et al. (2021a), and 3D human texture estimation Xu & Loy (2021). Due to the powerful global information modeling ability, these Transformer-based methods achieve significant performance gain compared with CNNs that focus on local information. In this paper, we successfully manage to take advantages of Vision Transformers for motion transfer with a warping and generation two branch architecture. The two branches employ cross-attention and convolution to enrich generation quality from both global and local viewpoints. Moreover, we propose a novel mutual learning loss, which regularizes the two branches to learn the advantages of each other, thus effectively increasing photorealism.

3 PROPOSED METHOD

Fig. 2 shows an overview of MotionFormer. It consists of two Transformer encoders and one Transformer decoder. The two Transformer encoders first extract the features of source image I_s and target pose image P_t, respectively. Then the Transformer decoder builds the relationship between I_s and P_t with two-branch decoder blocks hierarchically. Finally, a fusion block predicts the reconstructed person image I_{out}. The network is trained end-to-end with the proposed mutual learning loss. We utilize the Cross-Shaped Window Self-Attention (CSWin Attention) Dong et al. (2022) as our Attention mechanism in the encoder and decoder. The CSWin Attention calculates attention in the horizontal and vertical stripes in parallel to ensure the performance and efficiency. In our method, we assume a fixed background and simultaneously estimate a foreground mask M_{out} to merge an inpainted background with I_{out} in the testing phase. We introduce the Transformer encoder and decoder in Sec. 3.1 and Sec. 3.2 respectively. The mutual constrain loss is presented in Sec. 3.3.

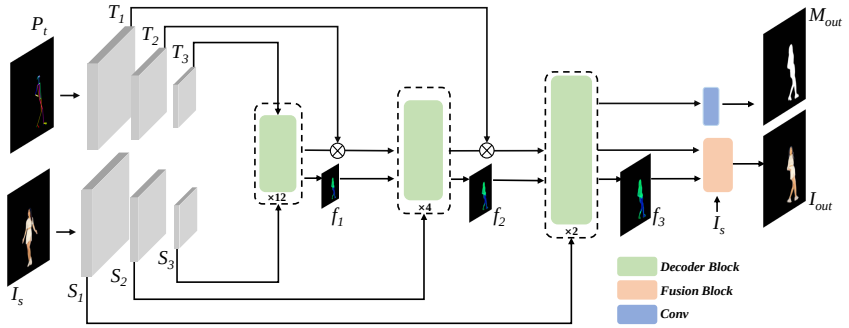


Figure 2: Overview of our MotionFormer framework. We use two Transformer encoders to extract features of the source image I_s and the target pose image P_t. These two features are hierarchically combined in one Transformer decoder where there are multiple decoder blocks. Finally, a fusion block synthesizes the output image by blending the warped and generated images.

3.1 Transformer Encoder

The structure of the two Transformer encoders is the same. Each encoder consists of three stages with different spatial size. Each stage has multiple encoder blocks and we adopt the CSwin Transformer Block Dong et al. (2022) as our encoder block. Our Transformer encoder captures hierarchical representations of the source image I_s and the target pose image P_t. S_i and T_i ($i = 1, 2, 3$) denote the output of the i-th stage for I_s and P_t, respectively, as shown in Fig. 2. We follow Dong et al. (2022); Wu et al. (2021) to utilize a convolutional layer between different stages for token reduction and channel increasing. We show more details of the encoder in the appendix.

3.2 Transformer Decoder

There are three stages in our Transformer decoder. The number of the decoder block in each stage is 2, 4, and 12, respectively. We concatenate the output of each stage and the corresponding target pose feature by skip-connections. The concatenated results are sent to the second and third stages. Similar to the encoder, we set the convolutional layer between different stages to increase token numbers and decrease channel dimensions.

3.2.1 Decoder block.

As shown in Fig 3, the decoder block has warping and generation branches. In each branch, there is a cross-attention process and a convolutional layer to capture the global and local correspondence respectively. Let X_{de}^{l} denote the output of l-th decoder block ($l > 1$) or the output of precedent stage ($l = 1$). For the first decoder stage, we set the T_3 as input so the $X_{de}^{1} = T_3$. The decoder block first extracts $\hat{X}_{de}^{l}$ from X_{de}^{l-1} with a Multi-Head Self-Attention process. Then we feed $\hat{X}_{de}^{l}$ to the warping branch and generation branch as Query (Q), and we use the feature of source encoder S_i as Key (K) and Value (V) to calculate the cross-attention map similar to Vaswani et al. (2017) with Multi-Head Cross-Attention process. The cross-attention map helps us build the global correspondence between the target pose and the source image. Finally, we send the output of the Multi-Head Cross-Attention to a convolutional layer to extract the local correspondence. The warping branch predicts a flow field to deform the source feature conditioned on the target pose, which helps the generation of regions that are visible in the source image. While for the invisible parts, the generation branch synthesizes novel content with the contextual information mined from the source feature. We combine the advantages of these two branches in each decoder block to improve the generation quality.

Warping branch. The warping branch aims to generate a flow field to warp the source feature S_i. Specifically, the Multi-Head Cross-Attention outputs the feature with the produced Q, K, V, and we feed the output to a convolution to inference the flow field. Inspired by recent approaches that gradually refine the estimation of optical flow Hui et al. (2018); Han et al. (2019b), we estimate a

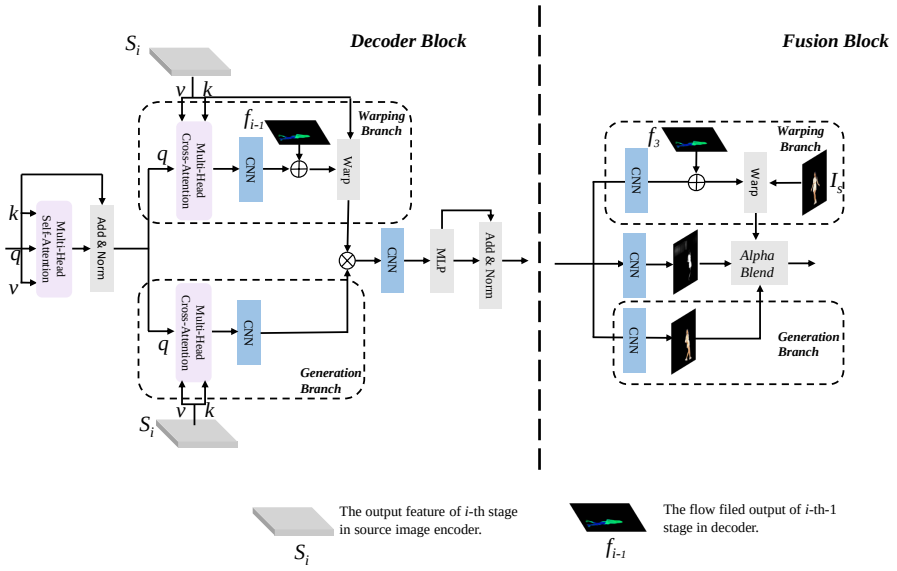


Figure 3: Overview of our decoder and fusion blocks. There are warping and generation branches in these two blocks. In decoder block, We build the global and local correspondence between source image and target pose with Multi-Head Cross-Attention and CNN respectively. The fusion block predict an mask to combine the output of two branches in pixel level.

residual flow to refine the estimation of the previous stage. Next, we warp the feature map S_i according to the flow field using bilinear interpolation. Formally, the whole process can be formulated as follows:

$$\begin{aligned} &Q = W^Q(\hat{X}^l_{de}), K = W^K(S_i), V = W^V(S_i), \\ &f^l = \text{Conv}(\text{Multi-Head Cross-Attention}(Q, K, V))), \\ &f^l = \text{Up}(f_{i-1}) + f^l, \text{if } l = 1 \text{ and } i > 1, \\ &O^l_w = \text{Warp}(S_i, f^l), \end{aligned} \tag{1}$$

where W^Q, W^K, W^V are the learnable projection heads in the self-attention module, the O^l_w denotes the output of the warping branch in l-th block, the Up is a $\times 2$ nearest-neighbor upsampling, and Warp denotes warping feature map S_i according to flow f^l using grid sampling Jaderberg et al. (2015). For the i-th decoder stage, the flow predicted by the last decoder block is treated as f_i and then refined by the subsequent blocks.

Generation branch. The architecture of the generation branch is similar to the warping branch. The attention outputs the feature with the produced Q, K, V, and then we feed the output to a convolution to infer the final prediction O^l_g:

$$\begin{aligned} &Q = W^Q(\hat{X}^l_{de}), K = W^K(S_i), V = W^V(S_i), \\ &O^l_g = \text{Conv}(\text{Multi-Head Cross-Attention}(Q, K, V))). \end{aligned} \tag{2}$$

where W^Q, W^K, W^V are the learnable projection heads in the self-attention module. The generation branch can generate novel content based on the global information of source feature S_i. Therefore, it is complementary to the warping branch when the flow field is inaccurate or there is no explicit reference in the source feature. Finally, we concatenate the output of warping and generation branch and reduce the dimension with a convolutional layer followed by an MLP and a residual connection:

$$\begin{aligned} &\bar{X}^l_{de} = \text{Conv}(\text{ Concat } (O^l_w, O^l_g)), \\ &X^l_{de} = \text{MLP}(\text{LN}(\hat{X}^l_{de})) + \bar{X}^l_{de}, \end{aligned} \tag{3}$$

where the $\bar{X}^l_{de}$ is the combination of warping and generation branches in the l-th decoder block.

3.2.2 FUSION BLOCK.

The fusion block takes the decoder output to predict the final result. The fusion block has a warping branch and generation branch at the pixel level. The warping branch refines the last decoder flow field f_3 and estimates a final flow f_f. And the generation branch synthesizes the RGB value I_f. At the same time, a fusion mask M_f is predicted to merge the output of these two branches:

$$\begin{aligned} f_f &= \text{Conv}(O_{de}) + \text{Up}(f_3), \\ M_f &= \text{Sigmoid}(\text{Conv}(O_{de})), \\ I_f &= \text{Tanh}(\text{Conv}(O_{de})), \\ I_{out} &= M_f \odot \text{Warp}(I_s, f_f) + (1 - M_f) \odot I_f, \end{aligned} \tag{4}$$

where the O_{de} is the output of decoder, $\odot$ is the element-wise multiplication, and I_{out} is the final prediction.

3.3 MUTUAL LEARNING LOSS

The generation and warping branches have their own advantages as mentioned above. Intuitively, we concatenate the output of these two branches followed by a convolution layer and an MLP as shown in Fig. 2, but we empirically find the convolution layer and MLP cannot combine these advantages well (see Sec. 5). To address this limitation and ensure the results have both advantages of these two branches, we propose a novel mutual learning loss to enforce these two branches to learn the advantages of each other. Specifically, the mutual learning loss enables these two branches to supervise each other within each decoder block, let $O_w^k, O_g^k \in R^{(H \times W) \times C}$ denote the reshaped outputs of the last warping and generation branch at the k-th decoder stage (see Eqs. (1) and (2) for their definition). If we calculate the similarity between the feature vector $O_{w,i}^k \in R^C$ at the spatial location i of O_w^k and all feature vectors $O_{g,j}^k \in R^C$ $(j = 1, 2, \ldots, HW)$ in O_g^k, we argue that the most similar vector to $O_{w,i}^k$ should be $O_{g,i}^k$, which is at the same position in O_g^k. In another word, we would like to enforce $i = \arg\max_j \text{Cos}(O_{w,i}^k, O_{g,j}^k)$, where $\text{Cos}(\cdot, \cdot)$ is the cosine similarity. This is achieved by the following mutual learning loss:

$$L_{\text{mut}} = \sum_k \sum_{i=1}^{HW} ||\underset{j}{\text{SoftArgMax}}(\text{Cos}(O_{w,i}^k, O_{g,j}^k)) - i||_1, \tag{5}$$

where the SoftArgMax is a differentiable version of $\arg\max$ that returns the spatial location of the maximum value. The mutual learning loss constrain the two branches to have high correlations at the same location, enhancing the complementariness of warping and generation.

In addition to the perceptual diversity loss, we follow the Ren et al. (2020) and Huang et al. (2021a) utilize the reconstruction loss Johnson et al. (2016), feature matching loss Wang et al. (2018c), hinge adversarial loss Lim & Ye (2017), style loss Gatys et al. (2015), total variation loss Johnson et al. (2016) and mask loss Huang et al. (2021a) to optimize our network. We show the details in the appendix.

4 EXPERIMENTS

Datasets. We use the solo dance YouTube videos collected by Huang et al. (2021a) and iPer Liu et al. (2019) datasets. These videos contain nearly static backgrounds and subjects that vary in gender, body shape, hairstyle, and clothes. All the frames are center cropped and resized to 256×256. We train a separate model on each dataset to fairly compare with other methods.

Implementation details. We use OpenPose Cao et al. (2017) to detect 25 body joints for each frame. These joints are then connected to create a target pose stick image P_t, which has 26 channels and each channel indicating one stick of the body. We use M_{gt} to separate foreground person image I_{gt} from the original frame. Our model is optimized using Adam optimizer with $\beta_1 = 0.0$, $\beta_2 = 0.99$, and initial learning rate of 10^{-4}. We utilize the TTUR strategy Heusel et al. (2017) to train our model. During the inference phase, we select one image from one video as the source image. The target background is generated by inpainting the source image background with LAMA Suvorov

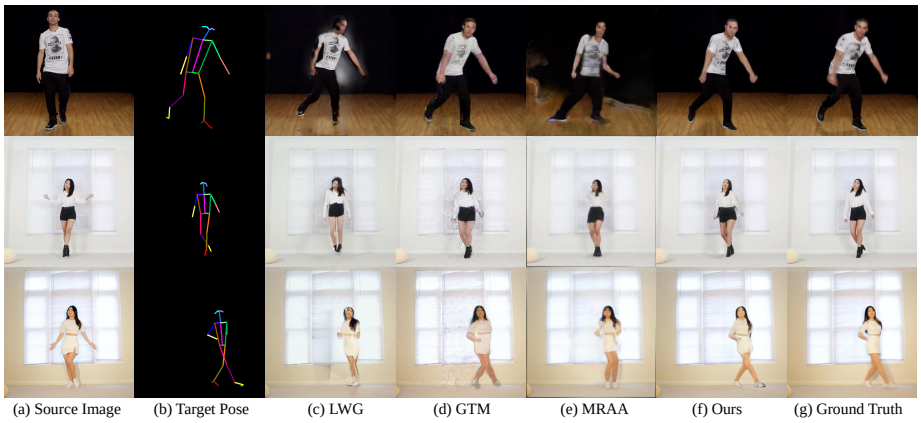


Figure 4: Visual comparison of state-of-the-art approaches and our method on YouTube videos dataset. Our proposed framework generates images with the highest visual quality.

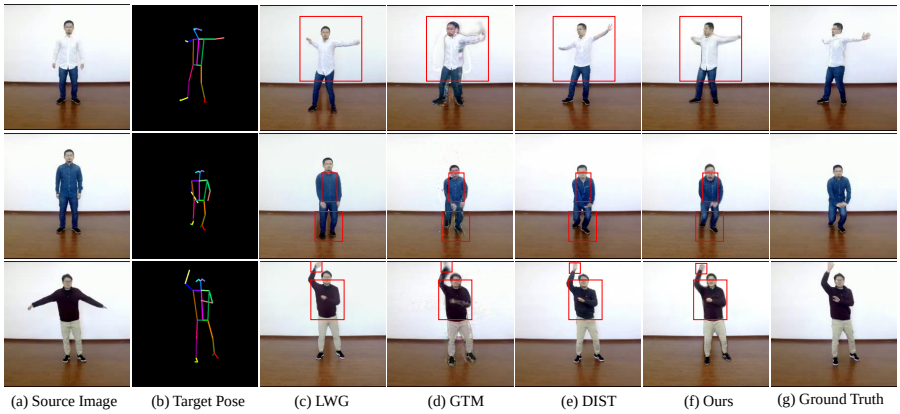


Figure 5: Visual comparison of state-of-the-art approaches and our method on iPer dataset. Our proposed framework generates images with the highest visual quality.

et al. (2021). When the target person has different body shapes (*e.g.*, heights, limb lengths) or the target person and source person are at different distances from the camera, we use the strategy in Chan et al. (2019) to normalize the pose of the target human.

Baselines. We compare MotionFormer with the state-of-the-art human motion transfer approaches: LWG Liu et al. (2019), GTM Huang et al. (2021a), MRAA Siarohin et al. (2021) and DIST Ren et al. (2020). For LWG, we test it on iPer dataset with the released pre-trained model, and we train LWG on YouTube videos with its source code. At test time, we fine-tune LWG on the source image as official implementation (fine-tuning is called "personalize" in the source code). For GTM, we utilize the pre-trained model on the YouTube videos dataset provided by the authors and retrain the model on iPer with the source code. As GTM supports testing with multiple source images, we use 20 frames in the source video and fine-tune the pre-trained network as described in the original paper Huang et al. (2021a). For MRAA, we use the source code provided by the authors to train the model. For DIST Ren et al. (2020), we compare with it using the pre-trained model on iPer dataset. For synthesizing a 1,000 frame video, the average per frame time costs of MRAA, LWG, GTM, DIST, and our method are 0.021s, 1.242s, 1.773s, 0.088s, and 0.94s, respectively. Meanwhile, MotionFormer does not require an online fine-tuning, while LWG and GTM do.

Method	PSNR↑	SSIM↑	LPIPS↓	FID↓	User study↓
LWG Liu et al. (2019)	18.94	0.686	0.175	85.06	97.61%
GTM Huang et al. (2021a)	21.50	0.819	0.137	77.69	76.19%
MRAA Siarohin et al. (2021)	18.95	0.674	0.234	160.97	100%
Ours	**23.50**	**0.885**	**0.073**	**65.03**	-

Table 1: Quantitative comparisons of state-of-the-art methods on YouTube videos dataset. User study denotes the preference rate of our method against the competing methods. Chance is 50%.

Method	PSNR↑	SSIM↑	LPIPS↓	FID↓	User study↓
LWG Liu et al. (2019)	23.93	0.843	0.089	40.41	58.90%
GTM Huang et al. (2021a)	23.94	0.840	0.120	60.26	75.00%
DIST Ren et al. (2020)	24.19	0.852	0.071	30.34	62.50%
Ours	**24.72**	**0.856**	**0.069**	**27.25**	-

Table 2: Quantitative comparisons of state-of-the-art methods on iPer videos dataset. User study denotes the preference rate of our method against the competing methods. Chance is 50%.

4.1 Qualitative Comparisons

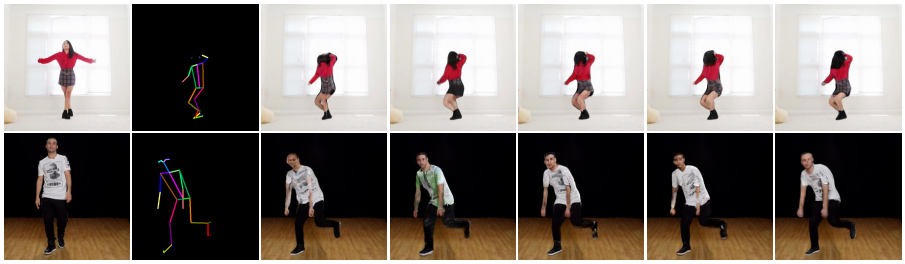
Qualitatively comparisons are given in Fig. 4 and Fig. 5. Although LWG Liu et al. (2019) can maintain the overall shape of the human body, it fails to reconstruct complicated human parts (*e.g.*, long hair and shoes in Fig. 4) of the source person and synthesis image with a large body motion (*e.g.*, squat in the red box of Fig. 4), which leads to visual artifacts and missing details. This is because LWG relies on the 3D mesh predicted by HMR Kanazawa et al. (2018), which is unable to model detailed shape information. In contrast, GTM Huang et al. (2021a) reconstructs better the body shape as it uses multiple inputs to optimize personalized geometry and texture. However, the geometry cannot handle the correspondence between the source image and the target pose. The synthesized texture also presents severe artifacts, especially for invisible regions in the source images. As an unsupervised method, MRAA Siarohin et al. (2021) implicitly models the relationship between source and target images. Without any prior information about the human body, MRAA generates unrealistic human images. DIST Huang et al. (2021a) does not model correct visual correspondence (*e.g.*, the coat buttons are missing in the last example) and suffers from overfitting (*e.g.*, the coat color becomes dark blue in the third example). Compared to existing methods, MotionFormer renders more realistic and natural images by effectively modeling long-range correspondence and local details.

4.2 Quantitative Comparisons

We use SSIM Wang et al. (2004), PSNR, FID Heusel et al. (2017), and LPIPS Zhang et al. (2018) as numerical evaluation metrics. The quantitative results are reported in Table 1 and Table 2. We observe that our method outperforms existing methods by large margins across all metrics. Additionally, we perform a human subjective evaluation. We generate the motion transfer videos of these different methods by randomly selecting 3-second video clips in the test set. On each trial, a volunteer is given compared results on the same video clip and is then asked to select the one with the highest generation quality. We tally the votes and show the statistics in the last column of Table 1 and Table 2. We can find that our method is favored in most of the trials.

5 Ablation Study

Attention Mechanism. To evaluate the effects of the cross-attention module, we delete the cross-attention in both the warping and generation branch. Instead, we concatenate the source feature S_i and Query directly in the Transformer decoder, followed by a convolution layer constructing their local relationship (this experiment is named Ours w/o Attention). As shown in Fig. 6(c), without modeling the long-range relationship between the source and target, Ours w/o Attention achieves worse results (*e.g.*, distorted skirt, limbs, and shoes). The numerical comparison shown in Table 3 is consistent with the visual observation.



(a) Source Image (b) Target Pose (c) w/o Attention (d) w/o warping (e) w/o generation (f) w/o mutual (g) Ours

Figure 6: Visual ablation study on YouTube videos dataset. (a) The source image. (b) The target pose. (c) Our method without Attention. (d) Our method without the warping branch. (e) Our method without the generation branch. (f) Our method without the mutual learning loss. (g) Our full method. Our full model can generate realistic appearance and correct body pose.

Generation and warping branches. We show the contributions of the generation branch and warping branch in the decoder block by removing them individually (*i.e.*, Ours w/o warping, Ours w/o generation). As shown in Fig. 6(d), without the warping branch, the generated clothing contains unnatural green and black regions in the man's T-shirt and woman's skirt, respectively. This phenomenon reflects that a single generation branch is prone to over-fitting. On the other hand, the warping branch can avoid over-fitting as shown in Fig. 6(e). However, the results still lack realism as the warping branch cannot generate novel appearances which are invisible in the source image (*e.g.*, the shoes of the man and the hair of the woman are incomplete). Our full model combines the advantages of these two branches and produces better results in Fig. 6(g). We also report the numerical results in Table 3, in which our full method achieves the best performance.

Mutual learning loss. We analyze the importance of mutual learning loss (Eq. (5)) by removing it during training (Ours w/o mutual). Fig. 6(f) shows the prediction combining the advantages of both warping and generation branches without using the mutual learning loss, which still produces noticeable visual artifacts. The proposed mutual learning loss aligns the output features from these two branches and improves the performance. The numerical evaluation in Table 3 also indicates that mutual learning loss improves the generated image quality. The other loss terms have been demonstrated effective in Balakrishnan et al. (2018); Han et al. (2019b); Wang et al. (2018b); Liu et al. (2019) with sufficient studies, so we do not include them in the ablation studies.

Method	PSNR↑	SSIM↑	LPIPS↓	FID↓
Ours w/o attention	20.95	0.842	0.194	81.10
Ours w/o warping	22.81	0.873	0.083	73.11
Ours w/o generation	22.39	0.872	0.100	70.21
Ours w/o mutual	22.73	0.876	0.078	68.70
Ours	**23.50**	**0.885**	**0.073**	**65.03**

Table 3: Ablation analysis of our proposed method on YouTube dataset. Our Full method achieves results that are superior to all other variants.

6 Conclusion

In this paper, we introduce MotionFormer, a Transformers-based framework for realistic human motion transfer. MotionFormer captures the global and local relationship between the source appearance and target pose with carefully designed Transformer-based decoder blocks, synthesizing promising results. At the core of each block lies a warping branch to deform the source feature and a generation branch to synthesize novel content. By minimizing a mutual learning loss, these two branches supervise each other to learn better representations and improve generation quality. Experiments on a dancing video dataset verify the effectiveness of MotionFormer.

7 ETHIC DISCUSSIONS

This work introduces a motion transfer method that can transfer motions from one subject to another. It may raise the potential ethical concern that malicious actions can be transferred to anyone (*e.g.*, celebrities). To prevent action retargeting on celebrities, we may insert a watermark to the human motion videos. The watermark contains the original motion source, which may differentiate the celebrity movement. Meanwhile, we can construct a celebrity set. We first conduct face recognition on the source person; if that person falls into this set, we will not perform human motion transfer.

8 REPRODUCIBILITY STATEMENT

The MotionFormer is trained for 10 epochs, and the learning rate decays linearly after the 5-th epoch. We provide the pseudo-code of the training process in Algorithm 1. We denote the Transformer encoder of the source images as En_s, the Transformer encoder of the target pose as En_t, the Transformer decoder as De, and the discriminator as D. We set the batchsize as 4 and step$_{max}$ is obtained by dividing the image numbers of dataset by batchsize. Meanwhile, we show the details of the model architecture and loss function in the appendix, this information is useful for the reproduce process.

Algorithm 1 Training Process

Require: A set of source images I_s, target pose images P_t, and person mask images M_{gt}.
  for Epoch = 1, 2, 3, ..., 10 **do**
    for Step = 1, 2, 3, ..., step$_{max}$ **do**
      Sample a batch of source images I_s, target pose P_t, and person mask M_{gt}.
      Get $S_1, S_2, S_3 = En_s(I_s), T_1, T_2, T_3 = En_t(P_t)$;
      Get $I_{out}, M_{out}, f_f = De([S_1, S_2, S_3], [T_1, T_2, T_3])$;
      Calculate the adversarial loss in Equation (11) in the main paper;
      Update D;
      Calculate the loss in Equation (14) in the main paper;
      Update De, En_s, En_t;
    end for
    if Epoch ≥ 5 **then**
      Update learning rate;
    end if
  end for

REFERENCES

Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7297–7306, 2018.

Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8340–8348, 2018.

Chenjie Cao, Yuxin Hong, Xiang Li, Chengrong Wang, Chengming Xu, XiangYang Xue, and Yanwei Fu. The image local autoregressive transformer. *arXiv preprint arXiv:2106.02514*, 2021a.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021b.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7291–7299, 2017.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pp. 213–229. Springer, 2020.

Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5933–5942, 2019.

Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. *arXiv preprint arXiv:2103.14899*, 2021a.

Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12299–12310, 2021b.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. Soft-gated warping-gan for pose-guided person image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 474–484, 2018.

Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12124–12134, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8857–8866, 2018.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12873–12883, 2021.

Oran Gafni, Oron Ashual, and Lior Wolf. Single-shot freestyle dance reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 882–891, 2021.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

Artur Grigorev, Artem Sevastopolsky, Alexander Vakhitov, and Victor Lempitsky. Coordinate-based texture inpainting for pose-guided human image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12135–12144, 2019.

Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10471–10480, 2019a.

Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10471–10480, 2019b.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6626–6637, 2017.

Zhichao Huang, Xintong Han, Jia Xu, and Tong Zhang. Few-shot human motion transfer by personalized geometry and texture modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2297–2306, 2021a.

Zilong Huang, Youcheng Ben, Guozhong Luo, Pei Cheng, Gang Yu, and Bin Fu. Shuffle transformer: Rethinking spatial shuffle for vision transformer. *arXiv preprint arXiv:2106.03650*, 2021b.

Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8981–8989, 2018.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1125–1134, 2017.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28:2017–2025, 2015.

Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two transformers can make one strong gan. *arXiv preprint arXiv:2102.07074*, 1(3), 2021.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7122–7131, 2018.

Kwonjoon Lee, Huiwen Chang, Lu Jiang, Han Zhang, Zhuowen Tu, and Ce Liu. Vitgan: Training gans with vision transformers. *arXiv preprint arXiv:2107.04589*, 2021.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1833–1844, 2021.

Jae Hyun Lim and Jong Chul Ye. Geometric gan. *arXiv preprint arXiv:1705.02894*, 2017.

Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5904–5913, 2019.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12009–12019, 2022.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.

Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 406–416, 2017.

Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 99–108, 2018.

Natalia Neverova, Riza Alp Guler, and Iasonas Kokkinos. Dense pose transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 123–138, 2018.

Jialun Peng, Dong Liu, Songcen Xu, and Houqiang Li. Generating diverse structure for image inpainting with hierarchical vq-vae. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10775–10784, June 2021.

Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. *arXiv preprint arXiv:1906.05909*, 2019.

Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H Li, and Ge Li. Deep image spatial transformation for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7690–7699, 2020.

Yurui Ren, Yubo Wu, Thomas H Li, Shan Liu, and Ge Li. Combining attention with flow for person image synthesis. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 3737–3745, 2021.

Aliaksandra Shysheya, Egor Zakharov, Kara-Ali Aliev, Renat Bashirov, Egor Burkov, Karim Iskakov, Aleksei Ivakhnenko, Yury Malkov, Igor Pasechnik, Dmitry Ulyanov, et al. Textured neural avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2387–2397, 2019.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2377–2386, 2019a.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32: 7137–7147, 2019b.

Aliaksandr Siarohin, Oliver Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *CVPR*, 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16519–16529, 2021.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. *arXiv preprint arXiv:2109.07161*, 2021.

Jiale Tao, Biao Wang, Borun Xu, Tiezheng Ge, Yuning Jiang, Wen Li, and Lixin Duan. Structure-aware motion transfer with deformable anchor model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3637–3646, 2022.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357. PMLR, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. *arXiv preprint arXiv:2103.14031*, 2021.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018a.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8798–8807, 2018b.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, pp. 8798–8807, 2018c.

Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Bryan Catanzaro, and Jan Kautz. Few-shot video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 5013–5024, 2019.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122*, 2021a.

Zhendong Wang, Xiaodong Cun, Jianmin Bao, and Jianzhuang Liu. Uformer: A general u-shaped transformer for image restoration. *arXiv preprint arXiv:2106.03106*, 2021b.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677*, 2020.

Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. *arXiv preprint arXiv:2103.15808*, 2021.

Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose Flow: Efficient online pose tracking. In *British Machine Vision Virtual Conference (BMVC)*, 2018.

Xiangyu Xu and Chen Change Loy. 3d human texture estimation from a single image with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13849–13858, 2021.

Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9459–9468, 2019.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3657–3666, 2022.

Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6881–6890, 2021.

A Appendix

A.1 Details of Attention Process and Encoder

As mentioned in the main paper, we utilize the Cross-Shaped Window Self-Attention (CSWin Attention) Dong et al. (2022) as our Attention mechanism in the encoder and decoder block. The CSwin Attention is based on the standard N heads attention of the original Transformer layer Vaswani et al. (2017). The main difference is that the CSWin Attention calculates attention in the horizontal and vertical stripes in parallel. For the attention in horizontal stripes at the n-th head, the query $Q \in R^{(H\times W)\times d}$, key $K \in R^{(H\times W)\times d}$, and value $V \in R^{(H\times W)\times d}$ are evenly split into M non-overlapping stripes of equal width sw (*i.e.*, $sw = H/M$). Then, it computes the standard attention $(\text{Softmax}(QK^T/\sqrt{d})V + B)$ separately for each stripe, where B is the locally-enhanced positional encoding defined in Dong et al. (2022). Meanwhile, it adopts the same operation on the vertical stripes. Finally, the CSWin Attention concatenates the output of the horizontal (H-Attention) and vertical (V-Attention) to predict final results:

$$\begin{aligned} &\text{Attention}(Q,K,V) = \text{Concat}(\text{head}_1,\ldots,\text{head}_\text{N}), \\ &\text{head}_n = \begin{cases} \text{H-Attention}\ (QW_n^Q, KW_n^K, VW_n^V) & n \le N/2 \\ \text{V-Attention}\ (QW_n^Q, KW_n^K, VW_n^V) & n > N/2 \end{cases}, \end{aligned} \tag{6}$$

where $W_n^Q, W_n^K, W_n^V \in R^{C\times d}, d = C/N$. Next, an MLP with GELU non-linearity between them is used for further feature transformations. The LayerNorm (LN) and residual connection are applied both before the Attention and MLP. The whole encoder block is then formulated as follows:

$$\begin{aligned} &Q = \text{LN}(X^{l-1}), K = \text{LN}(X^{l-1}), V = \text{LN}(X^{l-1}), \\ &\hat{X}^l = \text{Attention}(Q,K,V) + X^{l-1}, \\ &X^l = \text{MLP}(\text{LN}(\hat{X}^l)) + \hat{X}^l, \end{aligned} \tag{7}$$

where $X^l \in R^{(H\times W)\times C}$ denotes the output of the l-th encoder block ($l > 1$) or the precedent convolutional layer for the first encoder block of each stage ($l = 1$). We set numbers of encoder block to 1, 2, 21 for the three encoding stages. S_i and T_i ($i = 1, 2, 3$) denote the output of the final encoder block of the i-th stage for I_s and P_t, respectively, as shown in Fig. 2.

A.2 Details of Loss Function

In addition to the proposed mutual learning loss, we use losses below to optimize our model.

Reconstruction loss. We utilize the perceptual loss as the reconstruction loss, which minimize the feature map distance in an ImageNet-pretrained VGG-19 network Simonyan & Zisserman (2015):

$$L_{\text{rec}} = \sum_i \|F_i(I_{out}) - F_i(I_{gt})\|_1, \tag{8}$$

where F_i is the feature map of the i-th layer of the VGG-19 network. In our work, F_i corresponds to the activation maps from layers ReLU1_1, ReLU2_1, ReLU3_1, ReLU4_1, and ReLU5_1.

Feature matching loss. The feature matching loss L_{fm} compares the activation maps in the intermediate layers of the discriminator to stabilize training:

$$L_{\text{fm}} = \sum_i \|D_i(I_{out}) - D_{(i)}(I_{\text{gt}})\|_1, \tag{9}$$

where D_i is the activation of the i-th layer in the discriminator.

Style loss. We utilize the style loss to refine the texture of I_{out}. We write the style loss as:

$$L_{\text{style}} = \sum_i \frac{1}{M_i} \|G_i^F(I_{\text{out}}) - G_i^F(I_{\text{gt}})\|_1, \tag{10}$$

where G_i^F is a $C_i \times C_i$ Gram matrix computed given the feature map F_i, and M_i is the number of elements in G_i^F. These feature maps are the same as those used in the perceptual loss as illustrated above.

Hinge adversarial loss. We adopt the hinge adversarial loss to train our discriminator D and generator G:

$$\begin{aligned} L_{\text{adv}}^{\text{D}} &= -\mathbb{E}\left[h(D(I_{gt}))\right] - \mathbb{E}\left[h(-D(I_{out}))\right], \\ L_{\text{adv}}^{\text{G}} &= -\mathbb{E}\left[D(I_{out})\right], \end{aligned} \tag{11}$$

where $h(t) = min(0, -1 + t)$ is a hinge function used to regularize the discriminator.

Total variation loss. This loss regularizes the flow field f_f in the fusion block to be smooth, which is defined as

$$L_{\text{tv}} = \frac{1}{HW} \|\nabla f_f\|_1. \tag{12}$$

Mask loss. We feed the output of the decoder to a convolutional layer to predict a person mask M_{out}, which aims to combine the background and I_{out}n. We utilize the L_1 loss as the mask loss which can be written as:

$$L_{\text{mask}} = \sum_i \|M_{out} - M_{gt}\|_1, \tag{13}$$

where the M_{gt} is the ground-truth person mask, which is obtained with an off-the-shelf person segmentation model based on DeepLab V3 Chen et al. (2018).

Total losses. The total loss function can be written as:

$$\begin{aligned} L_{\text{total}} =& \lambda_{rec} \cdot L_{\text{re}} + \lambda_{fm} \cdot L_{\text{fm}} + \lambda_{adv} \cdot L_{\text{adv}}^{\text{G}} + \\ & \lambda_{mu} \cdot L_{\text{mu}} + \lambda_{tv} \cdot L_{\text{tv}} + \lambda_{mask} \cdot L_{\text{mask}} + \lambda_{style} L_{\text{style}}, \end{aligned} \tag{14}$$

where λ_{re}, λ_{fm}, λ_{adv}, λ_{mu}, λ_{tv}, λ_{mask}, and λ_{style} are the scalars controlling the influence of each loss term. Following the practice in Han et al. (2019b); Wang et al. (2018b), we set $\lambda_{re} = 10$, $\lambda_{fm} = 10$, $\lambda_{adv} = 1$, $\lambda_{mu} = 1$, $\lambda_{tv} = 0.5$, $\lambda_{mask} = 1$, and $\lambda_{style} = 10$.

A.3 Model Architecture

Fig. 7, Fig. 8 and Fig. 9 show the architectures of our Transformer encoder, decoder and discriminator in the training step. ic denotes the number of input channels and oc is the number of output channels. In the Transformer encoder, $ic = 26$ for the target pose and $ic = 3$ for the source person image. The $Flatten$ and $UnFlatten$ are reshaping operations to make the shape of features suitable for the attention and convolutional layers, respectively. The $Upsample$ operation doubles the spatial size and halves the number of channels of the input feature. The convolutional layers in the fusion block and mask prediction have the same architecture as this $Upsample$ operation but without the concatenation.

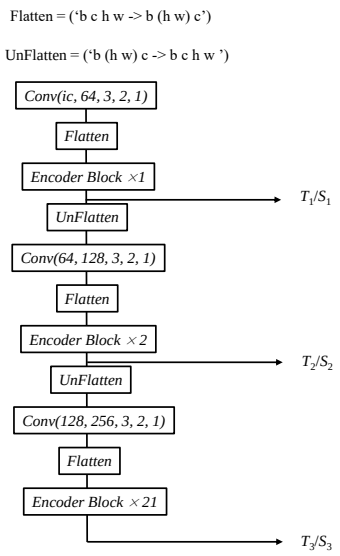


Figure 7: The detailed structure of our Transformer encoder. $Conv$ takes as input parameters of (the number of input channels, the number of output channels, filter size, stride, the size of zero padding).

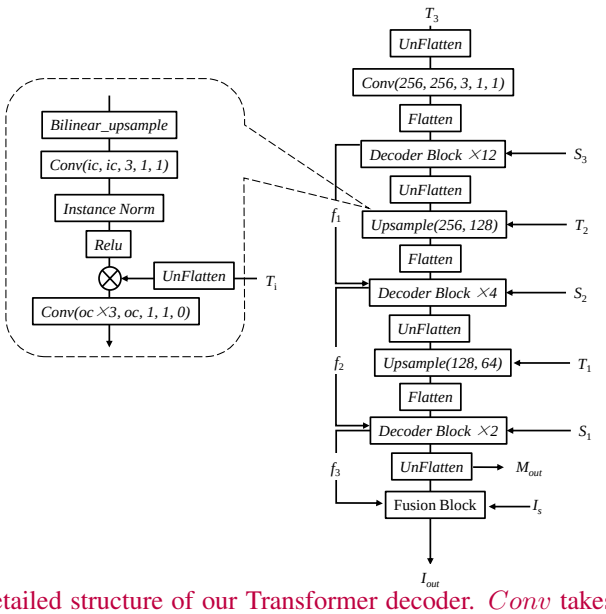


Figure 8: The detailed structure of our Transformer decoder. $Conv$ takes as input parameters of (the number of input channels, the number of output channels, filter size, stride, the size of zero padding).

B ANALYSIS OF NUMBER OF THE DECODER BLOCKS

We half the numbers of the decoder blocks to further analyze the performance of the decoder (this experiment is named Ours half). We set the number of each stage as 1, 2, and 6, respectively. As shown in Fig. 10, the results contain some artifacts when we reduce the number of decoder blocks. The numerical comparison shown in Table 4 is consistent with the visual observation. Please refer to the supplementary video for more results.

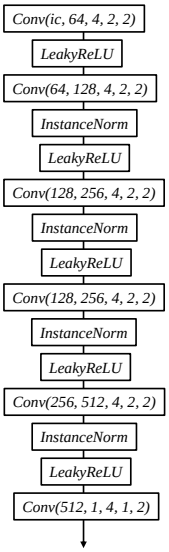
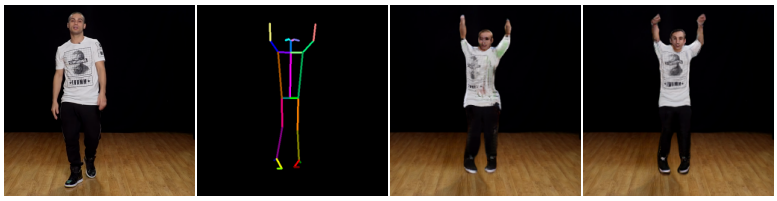


Figure 9: The detailed structure of our Discriminator. $Conv$ takes as input parameters of (the number of input channels, the number of output channels, filter size, stride, the size of zero padding).



(a) Source Image (b) Target Pose (c) half (d) Ours

Figure 10: Visual ablation study. (a) The source image. (b) The target pose. (c) Our method with half number of decoder blocks. (d) Our full method. Our full model can generate realistic appearance and correct body pose.

Method	PSNR↑	SSIM↑	LPIPS↓	FID↓
Ours half	22.56	0.876	0.092	71.20
Ours	**23.50**	**0.885**	**0.073**	**65.03**

Table 4: Ablation analysis of the number of decoder blocks.

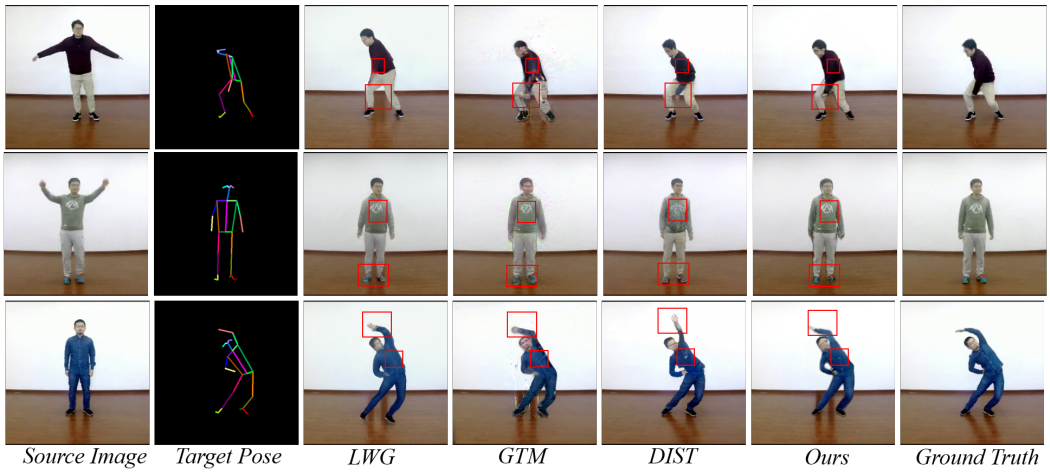


Figure 11: Visual comparison of state-of-the-art approaches and our method on iPer dataset. Our proposed framework generates images with the highest visual quality.

C More Visual Comparisons

We show more comparisons with LWG Liu et al. (2019), GTM Huang et al. (2021a), MRAA Siarohin et al. (2021), DIST Ren et al. (2020) in Fig. 12 and Fig. 11. The LWG fails to reconstruct the body with complicated human motion (*e.g.*, the squat in the red box of the first row in Fig. 11). In contrast, the GTM can synthesize the body shape better, but the textures are blurry. The MRAA captures the motion in an unsupervised manner, and the performance is constrained by the precision of motion prediction. DIST does not capture the correct relationship between the source image and target pose (*e.g.*, the arm in the last example in Fig. 11) and suffers from overfitting (*e.g.*, the shoes become red in the second example in Fig. 11). Overall, our method can effectively synthesize images with more accurate poses and finer texture details. We further provide video comparisons in the supplementary video. And we find our method can synthesize more visually pleasing and temporally coherent results.

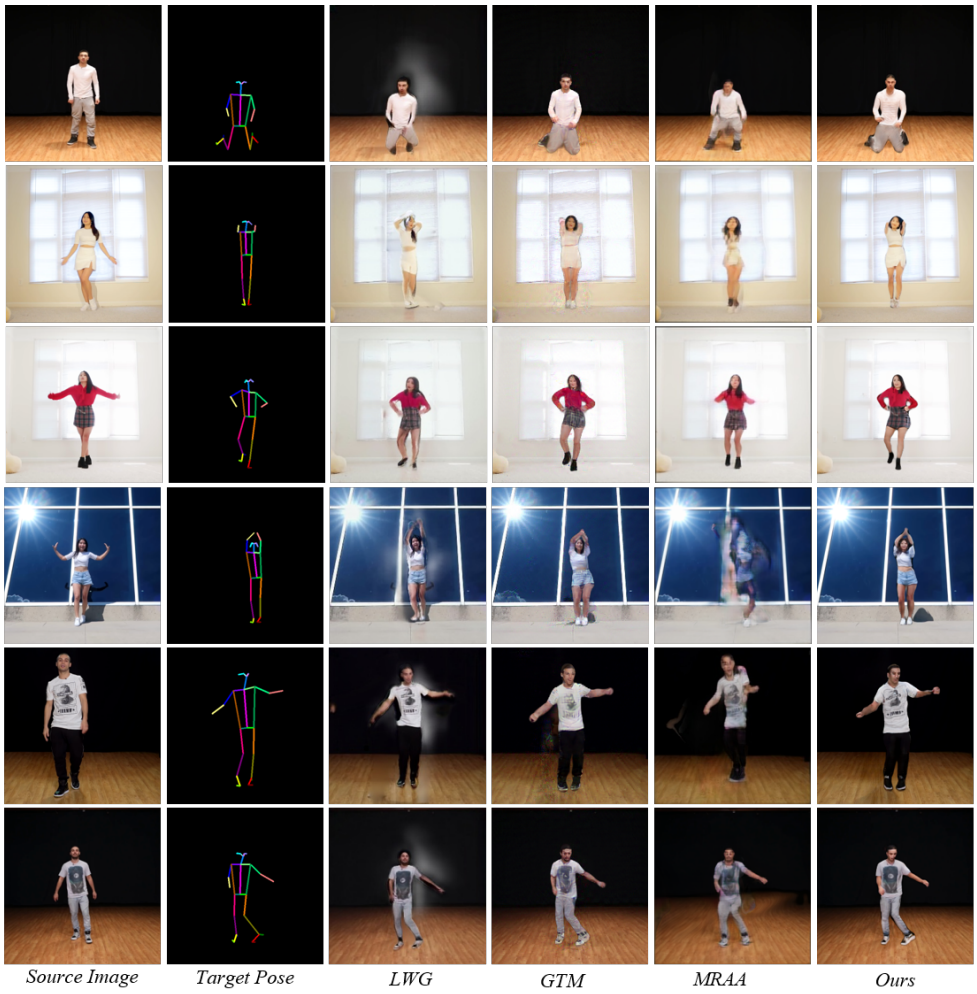


Figure 12: Visual comparison of state-of-the-art approaches and our method on YouTube dataset. Our proposed framework generates images with the highest visual quality.