HUMAN MOTIONFORMER: TRANSFERRING HUMAN MOTIONS WITH VISION TRANSFORMERS

Hongyu Liu[1*] **Xintong Han**[2*] **Chengbin Jin**[2] **Lihui Qian**[2] **Huawei Wei**[3] **Zhe Lin**[2]
Faqiang Wang[2] **Haoye Dong**[5] **Yibing Song**[4†] **Jia Xu**[2] **Qifeng Chen**[1†]
[1]Hong Kong University of Science and Technology [2]Huya Inc
[3]Tencent [4]AI[3] Institute, Fudan University [5] Carnegie Mellon University
hliudq@cse.ust.hk yibingsong.cv@gmail.com

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:

$$\begin{aligned} &\text{Attention}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_\text{N}), \\ &\text{head}_n = \begin{cases} \text{H-Attention}\ (QW_n^Q, KW_n^K, VW_n^V) & n \leq N/2 \\ \text{V-Attention}\ (QW_n^Q, KW_n^K, VW_n^V) & n > N/2 \end{cases}, \end{aligned} \tag{1}$$

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:

$$\begin{aligned} &Q = \text{LN}(X^{l-1}), K = \text{LN}(X^{l-1}), V = \text{LN}(X^{l-1}), \\ &\hat{X}^l = \text{Attention}(Q, K, V) + X^{l-1}, \\ &X^l = \text{MLP}(\text{LN}(\hat{X}^l)) + \hat{X}^l, \end{aligned} \tag{2}$$

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 pipeline.

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):

$$L_{\text{rec}} = \sum_i \|F_i(I_{out}) - F_i(I_{gt})\|_1, \tag{3}$$

[*]X. Han and H. Liu contribute equally. [†]Y. Song and Q. Chen are the corresponding authors.

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:

$$L_{\text{fm}} = \sum_i \|D_i(I_{out}) - D_{(i)}(I_{\text{gt}})\|_1, \tag{4}$$

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:

$$L_{\text{style}} = \sum_i \frac{1}{M_i} \|G_i^F(I_{\text{out}}) - G_i^F(I_{\text{gt}})\|_1, \tag{5}$$

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:

$$\begin{aligned} L_{\text{adv}}^{\text{D}} &= -\mathbb{E}\left[h(D(I_{gt}))\right] - \mathbb{E}\left[h(-D(I_{out}))\right], \\ L_{\text{adv}}^{\text{G}} &= -\mathbb{E}\left[D(I_{out})\right], \end{aligned} \tag{6}$$

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

$$L_{\text{tv}} = \frac{1}{HW} \|\nabla f_f\|_1. \tag{7}$$

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:

$$L_{\text{mask}} = \sum_i \|M_{out} - M_{gt}\|_1, \tag{8}$$

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:

$$\begin{aligned} L_{\text{total}} = &\lambda_{rec} \cdot L_{\text{re}} + \lambda_{fm} \cdot L_{\text{fm}} + \lambda_{adv} \cdot L_{\text{adv}}^{\text{G}} + \\ &\lambda_{mu} \cdot L_{\text{mu}} + \lambda_{tv} \cdot L_{\text{tv}} + \lambda_{mask} \cdot L_{\text{mask}} + \lambda_{style} L_{\text{style}}, \end{aligned} \tag{9}$$

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. (2019); Wang et al. (2018), 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. 1, Fig. 2 and Fig. 3 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.

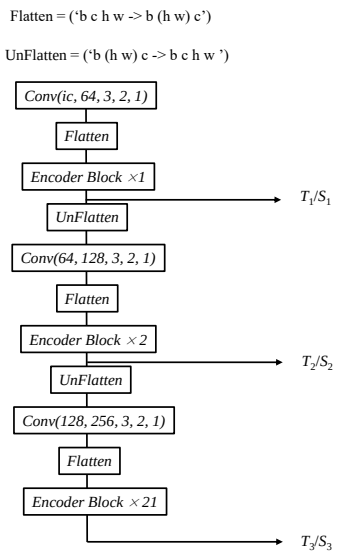


Figure 1: The detailed structure of our Transformer encoder. $Conv$ takes as input parameters of (the number of input channels, the number of output channels, filter size, stride, the size of zero padding).

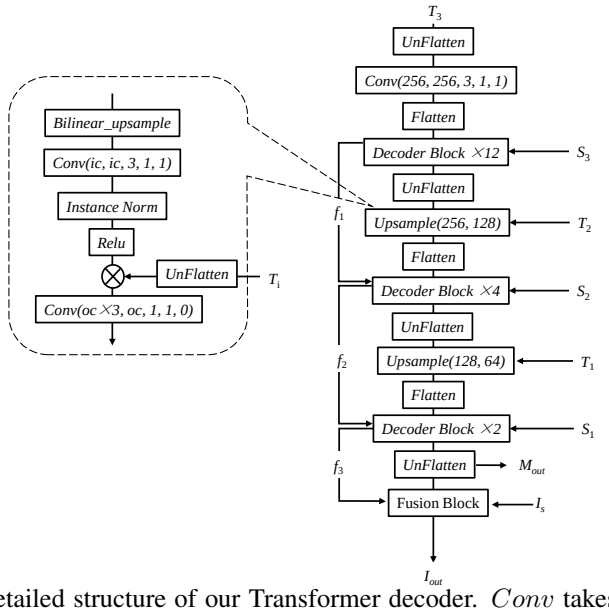


Figure 2: The detailed structure of our Transformer decoder. $Conv$ takes as input parameters of (the number of input channels, the number of output channels, filter size, stride, the size of zero padding).

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. 4, the results contain some artifacts when we reduce the number of decoder blocks. The numerical comparison shown in Table 1 is consistent with the visual observation. Please refer to the supplementary video for more results.

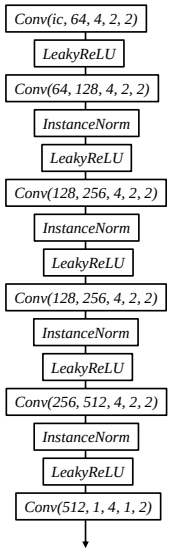
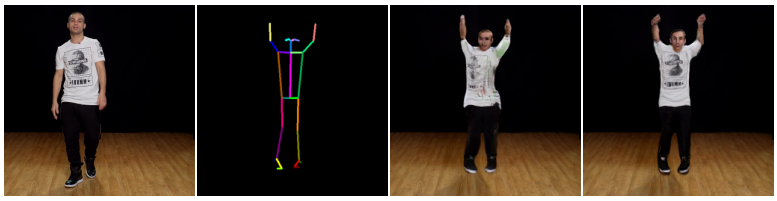


Figure 3: The detailed structure of our Discriminator. $Conv$ takes as input parameters of (the number of input channels, the number of output channels, filter size, stride, the size of zero padding).



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

Figure 4: 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 1: Ablation analysis of the number of decoder blocks.

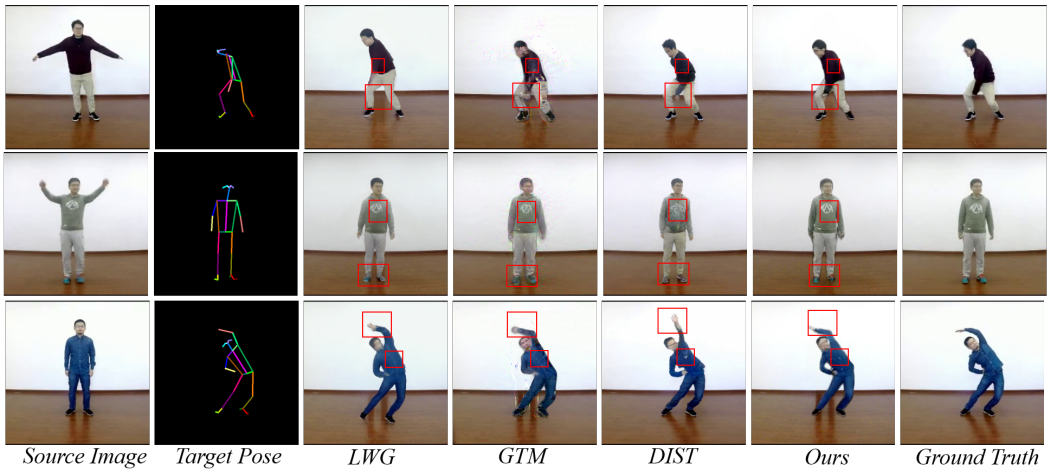


Figure 5: Visual comparison of state-of-the-art approaches and our method on iPer dataset. Our proposed framework generates images with the highest visual quality.

C More Visual Comparisons

We show more comparisons with LWG Liu et al. (2019), GTM Huang et al. (2021), MRAA Siarohin et al. (2021), DIST Ren et al. (2020) in Fig. 6 and Fig. 5. 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. 5). 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. 5) and suffers from overfitting (*e.g.*, the shoes become red in the second example in Fig. 5). 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.

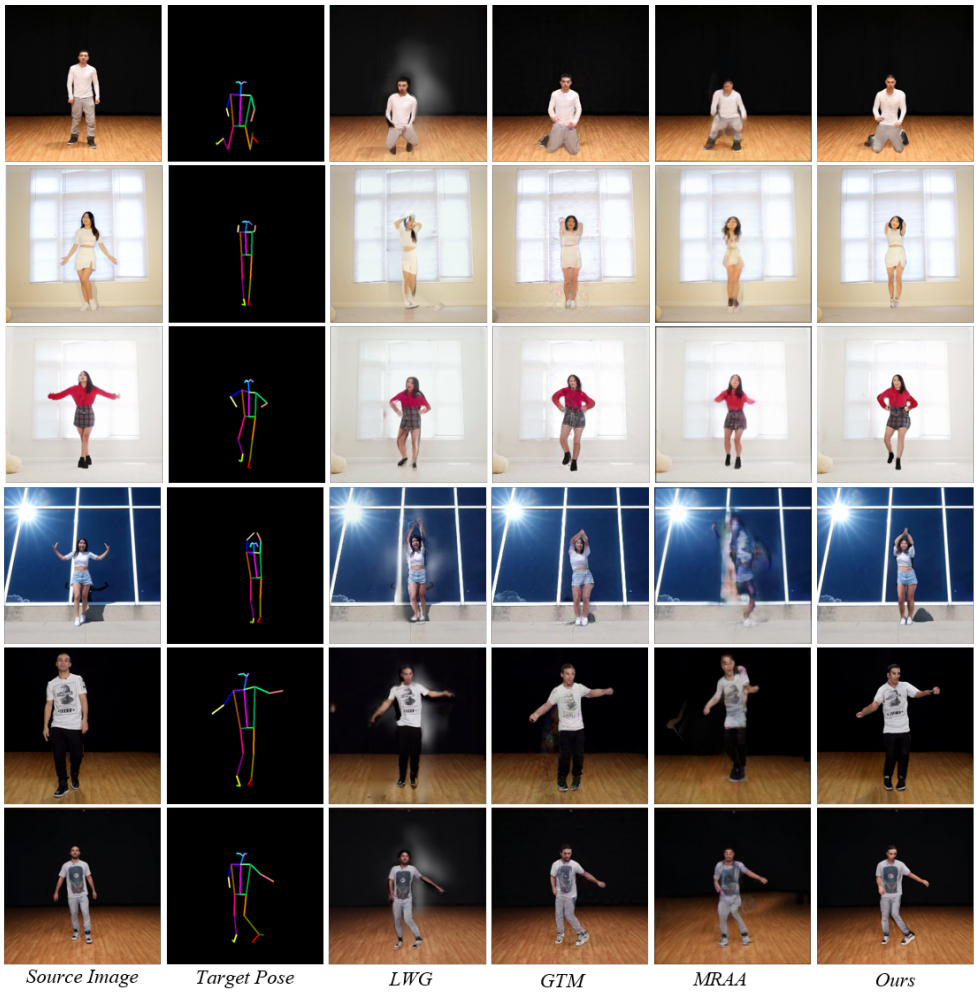


Figure 6: Visual comparison of state-of-the-art approaches and our method on YouTube dataset. Our proposed framework generates images with the highest visual quality.

REFERENCES

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.

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.

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, 2019.

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, 2021.

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.

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.

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.

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.

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, 2018.