Supplementary Material *for* Bidirectionally Deformable Motion Modulation For Video-based Human Pose Transfer

Wing-Yin Yu, Lai-Man Po, Ray C.C. Cheung, Yuzhi Zhao, Yu Xue, Kun Li
Department of Electrical Engineering, City University of Hong Kong, Hong Kong, China
{wingyinyu8, yzzhao2, yuxue22, kunli25}-c@my.cityu.edu.hk {eelmpo, r.cheung}cityu.edu.hk

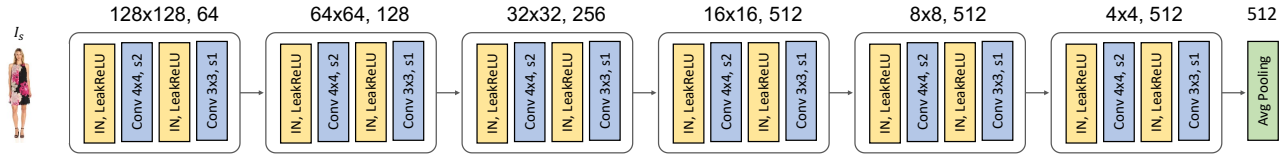

Figure 9. Network architecture of Style Encoder.

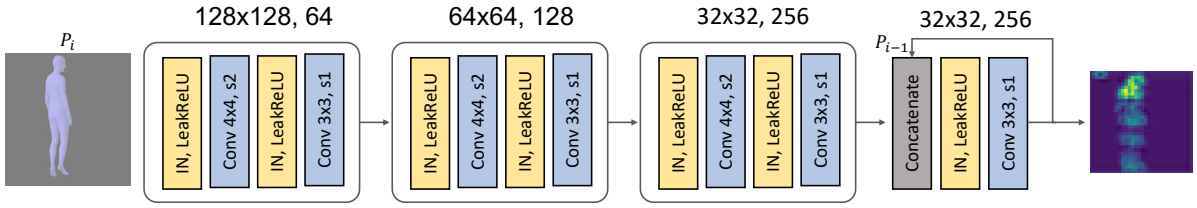

Figure 10. Network architecture of Structural Encoder.

1. Network Architecture

We present the design details of the proposed network architecture for different components. To be consistent, the resolution of all the input is 256×256. The input channels are set to 3 for a RGB image and 18 for a structural pose map. As depicted in Figure 9– 13, to simplify the notations, we use "IN" to represent Instance Normalization [19], "Conv $k \times k$, s#" to represent a convolutional layer with kernel size $k \times k$ and stride #. For example, "Conv 4×4, $s2$" indicates kernel size 4×4 and stride 2. With appropriate padding, we set the convolution layer with stride 2 to downscale the features to half of the input resolution.

1.1. Encoder

Style Encoder. The Style Encoder is designed to extract style code of the source image which is a vector that consists of dense semantic features from the source image. As shown in Figure 9, it includes 6 encoder blocks that progressively downsample the input features from 256×256 to 4×4. At the bottleneck of the encoder, we use an adaptive average pooling layer with kernel size 4×4 to compute the style vector.

Structural Encoder. The Structural Encoder is used to encode the spatial details of the target pose and shape so that it can produce a spatially aligned content in the final output image. Apart from cascaded convolutional blocks, we also leverage a recurrent flow at the bottleneck to maintain spatio-temporal information, as indicated in Figure 10. In this recurrent operation, we concatenate the input features and the output with same resolution. We visualize an example of output features in the Figure 10. The regions with key structural guidance such as eyes, hands or legs are well highlighted. The features with fading effect represent the hidden motion information. It indicates the effectiveness on extracting temporal information of the proposed recurrent Structural Encoder.

1.2. Discriminator

The discriminator is an essential element in our network to formulate the adversarial loss. As shown in Figure 11, there are two discriminators including the Spatial Discriminator and Temporal Discriminator in our framework. During training implementation, we randomly select a frame (4D tensor) from a mini-batch to calculate the spatial adversarial loss while using the whole mini-batch (5D tensor)

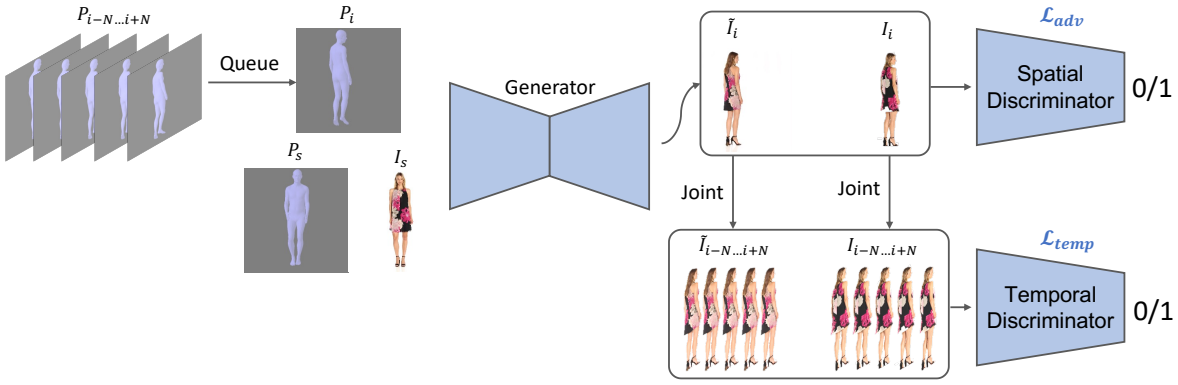

Figure 11. Overview of network architecture, including Spatial Discriminator and Temporal Discriminator.

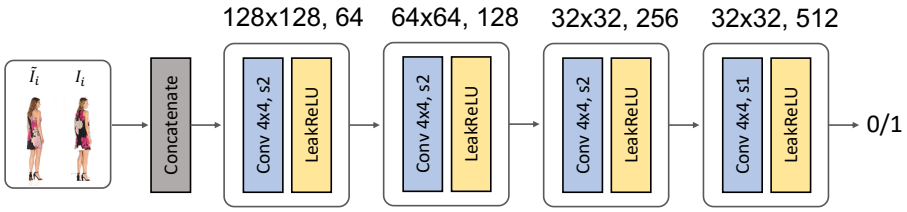

Figure 12. Network architecture of Spatial Discriminator.

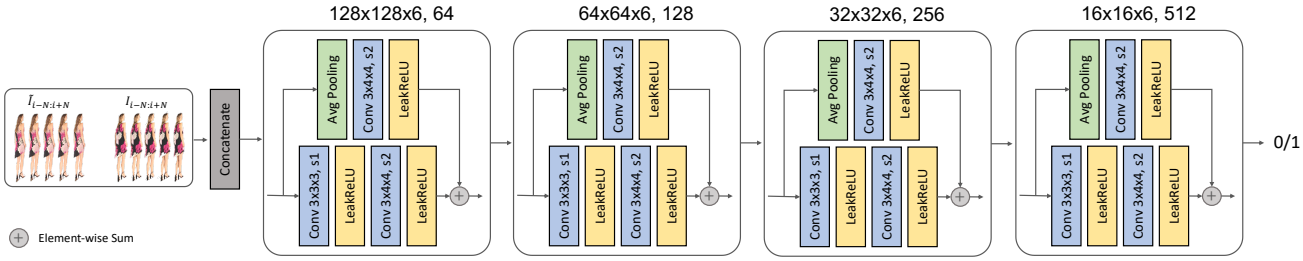

Figure 13. Network architecture of Temporal Discriminator.

to compute the temporal adversarial loss.

Spatial Discriminator. The Spatial Discriminator is used to mimic the distribution of the training set by discriminating whether the input pair is real or fake. We demonstrate the network architecture in Figure 12. Different from traditional GANs that using a single image as the input, we concatenate a generated image and a ground truth by channel dimension as a paired input, like PatchGAN [5]. There are 3 encoder blocks that progressively to reduce the resolution of input features from 256×256 to 32×32. We use a convolution layer with kernel size 4×4 and the LeaklyReLU [14] activation function to extract the patched features. Finally, we apply the least square error [13] to compute the statistical distance.

Temporal Discriminator. The temporal Discriminator is used to optimize the temporal consistency in time and feature channels of a mini-batch by using a 3D CNN model. During training, we collect the output image $\widetilde{I}_i$ one by one from time step $i - N$ to $i + N$. Similar with the Spatial Discriminator that concatenating the paired input images, we concatenate the generated sequence $\widetilde{I}_{i-N:i+N}$ and the target sequence $I_{i-N:i+N}$ by channel dimension as input. As indicated in Figure 13, there are 4 encoder blocks that progressively to downsample the input features from 256×256 to 16×16. The N is equal to 3 if the total iteration sequence length is 7. The major design of the encoder block is similar with one of Spatial Discriminator. We use 3D convolutional layer with kernel size $3 \times 3 \times 3$ or $3 \times 3 \times 4$ to downscale the input features by half. Moreover, we apply the average pooling layer to produce a downsampled residual map to preserve the feature signals. We use the weighted sum to fuse the output features and the residual branch.

Models	SSIM↑	PSNR↑	L1↓	FID↓	LPIPS↓	FVD-Train128f↓	FVD-Test128f↓
w/o L1 loss	N/A	15.147	0.117	379.232	0.388	2575.616±14.795	2626.936±29.864
w/o perceptual loss	0.914	23.767	0.0323	15.336	0.0518	174.383±2.413	159.217±6.818
w/o style loss	0.905	22.933	0.0373	14.956	0.0569	180.374±2.169	172.143±15.127
w/o CX loss	0.908	23.179	0.0349	14.345	0.0538	171.824±2.143	155.876±7.937
w/o spatial adv loss	0.913	23.576	0.0329	14.394	0.0506	201.065±3.101	187.118±8.642
w/o temporal adv loss	0.916	23.892	0.0312	14.466	0.0487	178.727±2.317	165.524±7.654
Full model	**0.918**	**24.071**	**0.0302**	**14.083**	**0.0478**	**168.275±2.564**	**148.253±6.781**

Table 3. Quantitative ablation study on the objective loss functions evaluated on the FashionVideo benchmark. The best scores are highlighted in **bold** format.

Figure 14. Pose transfer result on in-the-wild conditions.

2. Ablation Study on Loss Function

We conduct a comprehensive quantitative experiment on the analysis of objective loss functions in Table 3. We formulate the loss functions in three domains - pixel domain, semantic domain, and spatio-temporal domain to not only synthesize high-fidelity person image but also maintain details of person identity with characteristics of garments in the source image. The evaluation protocol is designed to observe the importance of each loss function by excluding the target function.

2.1. Pixel Domain

We mainly use L1 loss to minimize the absolute value of pixel distance between the generated image and the target image. The worst results on the model *w/o L1 loss* indicates the crucial role on generating acceptable images in our model. It is because pixel-wise comparison can preserve more global statistics such as basic appearance and shape of a person.

2.2. Semantic Domain

The losses on semantic domain are used to enhance the vividness of the generated images by comparing the features with different correspondence operations. It includes model *w/o perceptual loss*, model *w/o style loss*, and model *w/o CX loss*. The outcome shows that they all provide positive gains to the evaluation metrics in different aspects. The model *w/o perceptual loss* shows an 8% increment on the FID score. It represents the effectiveness on minimize the distribution distance between the generated results and the training set. The model *w/o style loss* and model *w/o CX loss* have major contributions on SSIM, PSNR, L1, and LPIPS scores. It indicates that these two losses can maintain more structural details on the generated images.

2.3. Spatio-temporal Domain

We mainly use adversarial losses to strengthen the spatio-temporal consistency of the generated sequence. The results of model *w/o spatial adv loss* and model *w/o temporal adv loss* demonstrates a large margin on the FVD-

Models	FLOPs
GFLA [18]	126.28G
Impersonator++ [12]	101.29G
DPTN [22]	30.97G
NTED [17]	103.99G
Ours	**16.92G**

Table 4. FLOPs comparison of the state-of-the-arts. The best scores are highlighted in **bold** format.

train128f and FVD-test128f scores compared to the final model. It can certify that these two losses can maintain spatio-temporally coherent information in our framework.

3. Mesh Flow Computation

Following previous work [9, 12], we demonstrate the procedures to obtain a mesh flow. Based on the $SMPL$ model, we can obtain the source weak perspective camera $K \in \mathbb{R}^{3\times1}$, explicit representation $V \in \mathbb{R}^{6890\times3}$, rasterized face function $H \in \mathbb{R}^{13776\times3}$, the barycentric weight index map of the triangulated face $W \in \mathbb{R}^{H\times W\times3}$, the 2D projected face index map of source mesh $C_s \in \mathbb{R}^{H\times W\times1}$ and target mesh C_i. We project the explicit vertices into a 2D coordinate system and get the triangulated faces $\hat{V} \in \mathbb{R}^{13776\times3\times2} = P(V, K, W)$. By matching the face index map of source mesh C_s and target mesh C_i, we can get the visible face index vector $Q \in \mathbb{R}^{13776}$. Finally, we get the $F_{i\to s}$ by multiplying the correspondence of source W with the visible triangulated faces $\hat{V}$, i.e.

$$F_{i\to s} = W \times (Q \times \hat{V}), \tag{14}$$

where $\times$ indicates matrix multiplication operation.

4. Computation Analysis

We further analyze the computational cost in term of FLOPs. As indicated in Table 4, our method just produces half as many FLOPs as the previous optimal SOTAs.

5. Transferring In-the-wild Images

We provide some results conditioned on some in-the-wild images in Figure 14 to demonstrate the generality. The resolution (256×256) is consistent with the evaluation datasets [11, 21]. The source images are randomly chosen from the Internet.

6. Limitation

With the success of bidirectional deformable modulation, it can synthesize a spatio-temporal video based on a set of noisy poses. However, the reconstruction of missing background is limited. It is believed that the deformable motion modulation is biased to capture the motion. For static background, the capability of motion offset and activation unit are not directly co-related. It could fail to recover some complex natural backgrounds such as trees.

7. Rationale behind DMM

We further elaborate the rationale behind the deformable demodulation. The motivation of weight demodulation is to base normalization on the expected statistics of the incoming style code without using instance normalization. Because instance normalization considers the average statistics of the instance-wise style features, but it produces significant artifacts when scaling the image as mentioned in StyleGANv2. Eqs 3 aims to scale the convolution weights according to the incoming style code so that it can combine the style statistics with the kernels. Eqs 4 proposes to normalize the weights by rescaling the L2 norm of itself so that it can restore the output features back to unit standard deviation.

8. Ablation Study on Hyperparameter

We conduct some ablation experiments on the hyperparameters of loss functions in Table 5. The evaluation protocol is designed to observe the importance of each loss function by fine-tuning the hypermeters.

9. Related Works

9.1. Diffusion Model

Recently, the diffusion model gains stunning generative performance on image synthesis field. Dhariwal *et. al.* [2] proposed a guided diffusion model that can extend the conditions from noisy to some structural signals. Ju *et. al.* [6] proposed a skeleton-guided diffusion model the transfer the style of source image to a key-point-based maps. However, the characteristics and identities of the source image could not be preserved in the generated images. The UPGPT [1] suggested a diffusion model that making use of textural description to embed the texture information. However, due to the limited generalization of text description, it is hard to formulate a precise text embedding for some textures with unique patterns such as labels or logos. DreamPose [7] propose an image-to-video diffusion model that using stable diffusion [4] to fuse CLIP encoder [16] and variational autoencoder to encode a source image. However, the lack of temporal consistency conditioned on noisy poses is still a challenging problem in this model. We fully appreciate the beauty of diffusion model that sparking the field of image synthesis recently. Although diffusion model can produce higher diversity image compared to GAN-based method, it is the smallest factor to drive video-based human pose

Models	SSIM↑	PSNR↑	L1↓	FID↓	LPIPS↓	FVD-Train128f↓	FVD-Test128f↓
$\lambda_{l1} = 0.5$	0.911	23.466	0.0336	16.355	0.0530	176.624±2.640	152.437±6.616
$\lambda_{l1} = 5$	0.913	23.737	0.0326	14.619	0.0500	180.083 ±2.175	152.407±7.776
$\lambda_{temp} = 1$	0.911	23.446	0.0334	15.604	0.0523	180.883±2.265	157.166±6.393
$\lambda_{temp} = 10$	0.913	23.526	0.0334	16.499	0.0531	176.185±2.340	157.942±6.180
$\lambda_{per} = 100$	0.912	23.762	0.0328	15.280	0.0522	177.532±2.236	156.503±6.426
$\lambda_{per} = 1000$	0.910	23.505	0.0344	15.519	0.0520	186.654±3.157	175.013±7.975
$\lambda_{gram} = 0.1$	0.911	23.293	0.0346	15.864	0.0532	177.645±1.722	157.355±7.569
$\lambda_{gram} = 1$	0.914	23.650	0.0330	15.504	0.0520	188.588±2.429	161.876±7.981
$\lambda_{cx} = 0.5$	Nan	15.147	0.1175	379.233	0.1175	2576.074±18.307	2628.241±31.276
$\lambda_{cx} = 1$	0.912	23.501	0.0336	14.938	**0.0336**	185.161±2.847	167.907±7.804
Full model	**0.918**	**24.071**	**0.0302**	**14.083**	0.0478	**168.275±2.564**	**148.253±6.781**

Table 5. Quantitative ablation study on the hyper parameters of objective loss functions evaluated on the FashionVideo benchmark. The best scores are highlighted in **bold** format.

transfer to do better because it requires the output should be appearance-aligned with the source image. Our GAN-based solution still has its merits in term of fast sampling and comparatively high-quality synthesis. Moreover, our main contributions focus more on improving the spatiotemporal consistency by smoothing the jittering poses. They are intuitive, effective and easily applied to other methods to enhance the performance as well.

9.2. Neural Rendering

Some related work to neural rendering for human appearance transfer and reenacting are also interesting. Instead of deformable human body models like SMPL, Prokudin *et. al.* [15] proposed SMPLpix to rasterize a sparse set of 3D mesh vertices into photorealistic images instead of using computer graphic engine. Gomes *et. al.* [3] proposed a shape-aware retargeting method based on a hybrid image-based rendering technique to perform human motion transfer. Liu *et. al.* [10] proposed to combine two convolutional neural networks to disentangle the learning of time-coherent information from the embedding of the human in 2D space. Kwon *et. al.* [8] introduced a Neural human performer that can learn a generalizable radiance field by a temporal transformer to aggregate tracked visual features. Xu *et. al.* [20] extended the uniform occupancy prior of traditional neural radiance field to a structured implicit human body model so that it can use signed distance functions. The effort of the neural rendering for human appearance transfer is highly appreciated. Our method and task are different from the neural rendering methods that restricted to one model per actor (subject-specific). They focus more on novel view synthesis, 3D reconstruction, and texture map rendering.

10. Video Comparison for SOTAs

We randomly select some video demonstrations to compare the visual quality with some state-of-the-art methods. The videos are from FashionVideo [21] and iPER [11] benchmarks. Please find the attached video_supplementary_iccv.zip file to enjoy the video clips. The default frame rate is 30fps.

References

[1] Soon Yau Cheong, Armin Mustafa, and Andrew Gilbert. Upgpt: Universal diffusion model for person image generation, editing and pose transfer. *arXiv preprint arXiv:2304.08870*, 2023. 4

[2] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 4

[3] Thiago L Gomes, Renato Martins, João Ferreira, Rafael Azevedo, Guilherme Torres, and Erickson R Nascimento. A shape-aware retargeting approach to transfer human motion and appearance in monocular videos. *International Journal of Computer Vision*, 129(7):2057–2075, 2021. 5

[4] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 4

[5] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[6] Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. Humansd: A native skeleton-guided diffusion model for human image generation. *arXiv preprint arXiv:2304.04269*, 2023. 4

[7] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. *arXiv preprint arXiv:2304.06025*, 2023. 4

[8] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Advances in Neural Information Processing Systems*, 34:24741–24752, 2021. 5

[9] Yining Li, Chen Huang, and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3693–3702, 2019. 4

[10] Lingjie Liu, Weipeng Xu, Marc Habermann, Michael Zollhöfer, Florian Bernard, Hyeongwoo Kim, Wenping Wang, and Christian Theobalt. Neural human video rendering by learning dynamic textures and rendering-to-video translation. *IEEE Transactions on Visualization and Computer Graphics*, PP:1–1, 05 2020. 5

[11] 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*, pages 5904–5913, 2019. 4, 5

[12] Wen Liu, Zhixin Piao, Zhi Tu, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan with attention: A unified framework for human image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 4

[13] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017. 2

[14] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010. 2

[15] Sergey Prokudin, Michael J Black, and Javier Romero. Smplpix: Neural avatars from 3d human models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1810–1819, 2021. 5

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[17] Yurui Ren, Xiaoqing Fan, Ge Li, Shan Liu, and Thomas H Li. Neural texture extraction and distribution for controllable person image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13535–13544, 2022. 4

[18] Yurui Ren, Ge Li, Shan Liu, and Thomas H Li. Deep spatial transformation for pose-guided person image generation and animation. *IEEE Transactions on Image Processing*, 29:8622–8635, 2020. 4

[19] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 1

[20] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *Advances in Neural Information Processing Systems*, 34:14955–14966, 2021. 5

[21] Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. Dwnet: Dense warp-based network for pose-guided human video generation. *arXiv preprint arXiv:1910.09139*, 2019. 4, 5

[22] Pengze Zhang, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Exploring dual-task correlation for pose guided person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7713–7722, 2022. 4