

• Supplementary File •

Weakly Supervised 2D Human Pose Transfer

Qian ZHENG^{†1}, Yajie LIU^{†1}, Zhizhao LIN¹,
Dani LISCHINSKI², Daniel COHEN-OR¹ & Hui HUANG^{*1}

¹Shenzhen University, Shenzhen 518060, China;

²The Hebrew University of Jerusalem, Jerusalem 91905, Israel

Appendix A Datasets

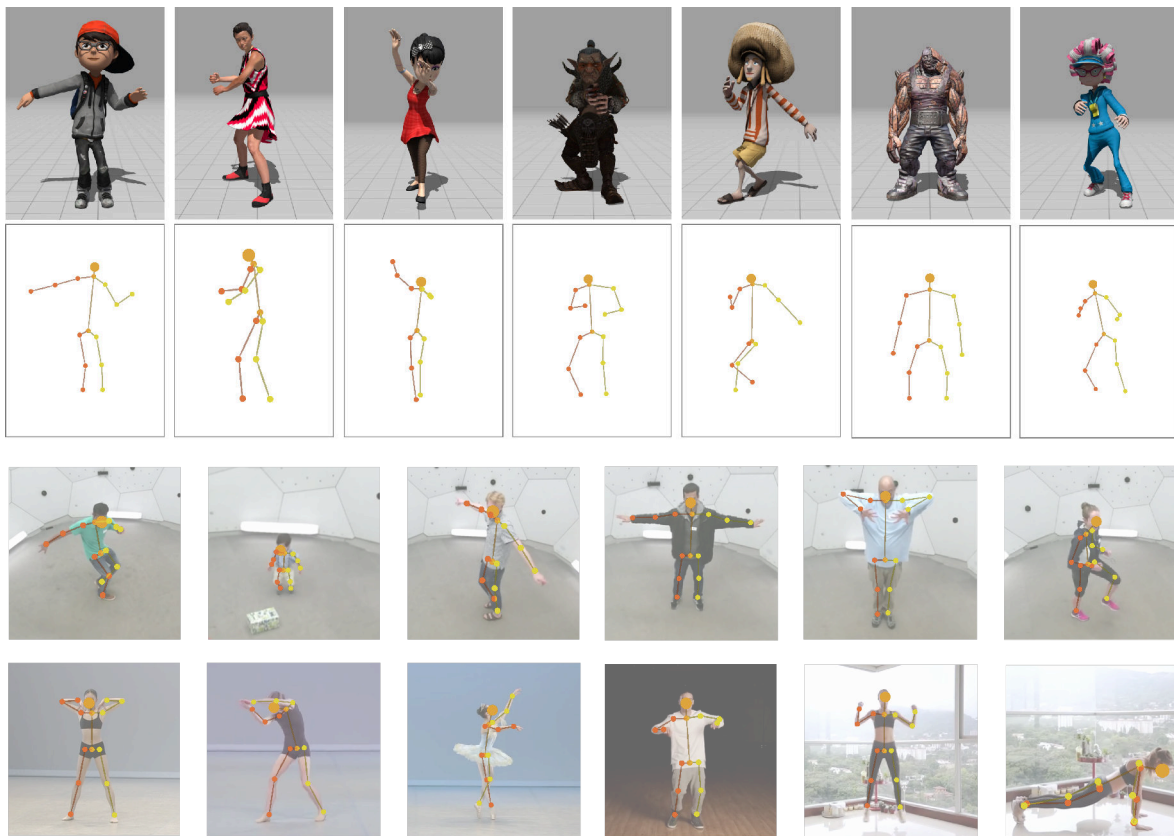


Figure A1 We train our network using 2D projections of 3D skeletons from two different motion datasets, together with 2D poses extracted from videos of people performing various actions. The first two rows show some pose samples from the Adobe Mixamo dataset. The third row presents samples from the CMU Panoptic Dataset. Six sample skeletons extracted from 2D videos are shown at the bottom row.

We train our network using 2D projections of 3D skeletons from two different motion datasets (the Adobe Mixamo dataset [1] and the CMU Panoptic Dataset [2]), together with 2D poses extracted from videos of people performing various actions. Figure A1 shows some sample poses, together with the images of corresponding characters or real persons. As can be seen, the training set contains the skeletons with different bone proportions.

Appendix B 3D Extension

We could extend our method to 3D pose transfer with two small modifications. The first one is representing the inputs and output by the collection of 3D positions of all joints. The second is considering three directions (X, Y, and Z axes) for virtual link loss.

* Corresponding author (email: hhzhiyan@gmail.com)

† Equal contribution

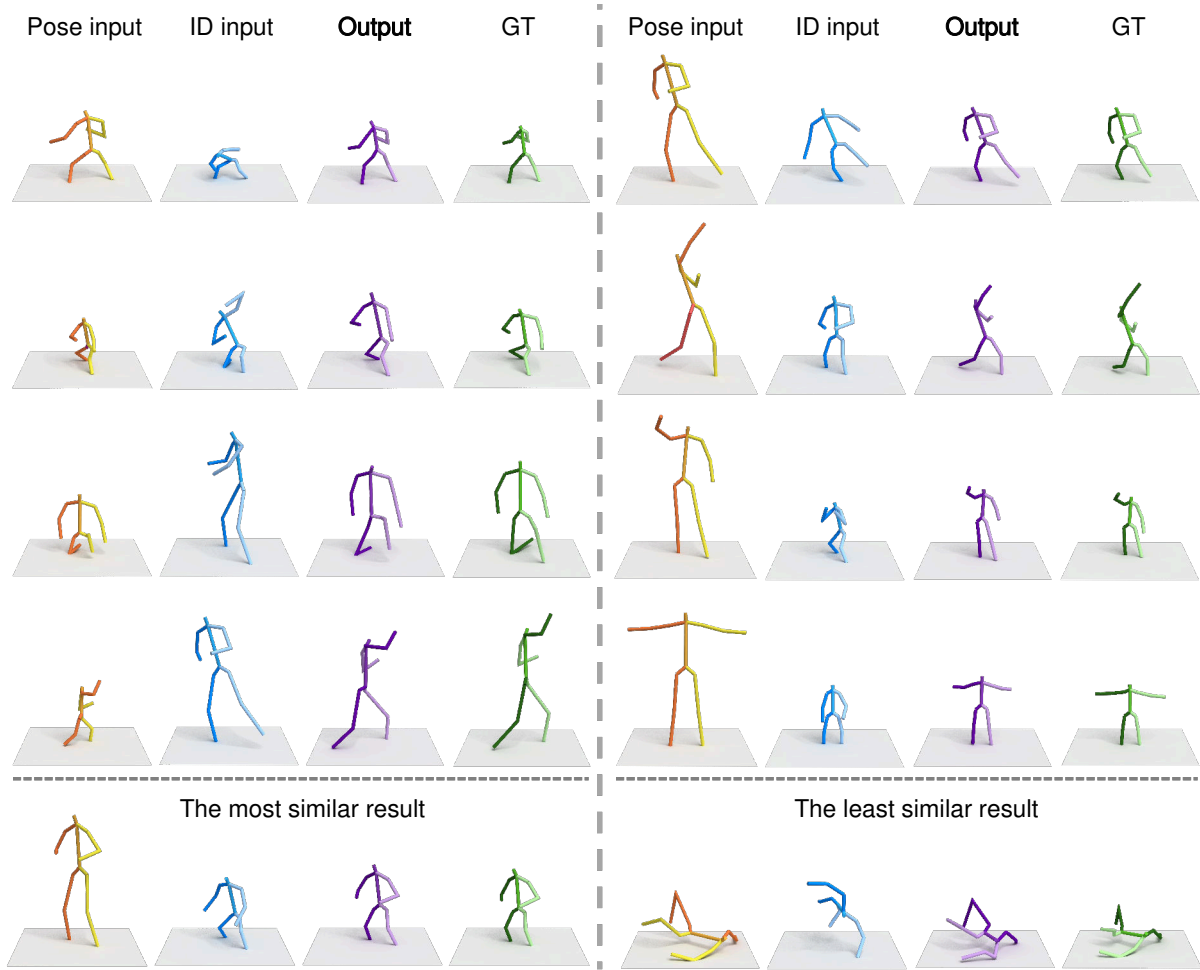


Figure B1 Our method extended to 3D skeletons: the first four rows present eight results that was randomly selected. The bottom row shows the transferred outputs with the smallest (left) and largest (right) errors computed with respect to the ID inputs, i.e., directly comparing the bone lengths between them.

Note that in the 3D case, though we could easily add a new loss term by directly comparing the bone lengths of the ID input and the output, we don't add it for consistency with 2D cases. We train the model using the same training set of Mixamo (without poses extracted from videos) and test it on the testing set. Figure B1 shows some results. By visual comparison, the generated poses are quite close to the ground truths.

For 3D skeletons, the output deviation from target ID can be accurately measured by comparing the bone lengths between them. Hence we define the error of output as $\sqrt{\sum_{j=1}^{14} (l_j^{\text{ID}} - l_j)^2 / 14}$, where l_j and l_j^{ID} denote the lengths of j -th bone in the output and ID input, respectively. The number of bones is 14. We measure the errors on 1M generated poses. The average error is 0.043, with a minimum of 0.01 and a maximum of 0.147; see the bottom row of Fig. B1.

Appendix C Evaluation on designed poses

Accurate frame-to-frame mapping is hard to obtain even different actors are performing similar motions. To get more accurate ground-truth for evaluation, we asked a professional 3D modeler to design six different and extreme (difficult) poses performed by three 3D characters (IDs) from the training set, such as picking up a ball or doing yoga.

These six poses performed by an ID are transferred to the other two IDs (ID-1 and ID-2). In Fig. C1, we show the transferred results, as well as the MSE (the lower, the better) and OKS (the higher, the better) compared to the ground truths. Note that ID-1 has a long upper body, ID-2 is the shortest, but the ID of pose input is tall and has long legs. Still, our generated poses preserve these ID features quite well. The results with the highest errors are highlighted in bold, which is a very challenging case. To make the hand tightly touch with the foot, the modeler sharply modified the directions of some bones. Such semantic information is hard to capture.

Appendix D More application examples

Figure D1 presents our network architecture for the layout generation application. It is with a dual-path encoder and a decoder, and the U-net structure [3] is used in both encoder and decoder. It takes the source pose heat map, the source cloth layout, the source pose mask, and the target pose heat map as input, and synthesizes the target layout. The cloth layout is represented by a 19-channel segmentation, and the pose heat map is represented by 18 channels, where each channel represents a joint. The pose

































Results Pose	ID		Evaluations		ID-2		Evaluations	
								
			OKS	67.05%			OKS	75.14%
			MSE of lengths	0.0016			MSE of lengths	0.0011
			MSE of angles	0.1059			MSE of angles	0.0862
			MSE of joints	0.0765			MSE of joints	0.0440
			OKS	54.80%			OKS	67.19%
			MSE of lengths	0.0023			MSE of lengths	0.0010
			MSE of angles	0.2099			MSE of angles	0.2416
			MSE of joints	0.1090			MSE of joints	0.0619
			OKS	81.95%			OKS	88.81%
			MSE of lengths	0.0013			MSE of lengths	0.0016
			MSE of angles	0.6047			MSE of angles	0.3012
			MSE of joints	0.0633			MSE of joints	0.0416
			OKS	71.88%			OKS	58.69%
			MSE of lengths	0.0017			MSE of lengths	0.0015
			MSE of angles	0.0113			MSE of angles	0.0182
			MSE of joints	0.0561			MSE of joints	0.0567
			OKS	68.16%			OKS	63.60%
			MSE of lengths	0.0009			MSE of lengths	0.0012
			MSE of angles	0.1118			MSE of angles	0.0961
			MSE of joints	0.0636			MSE of joints	0.0551
			OKS	58.84%			OKS	55.40%
			MSE of lengths	0.0014			MSE of lengths	0.0017
			MSE of angles	0.0497			MSE of angles	0.0209
			MSE of joints	0.0671			MSE of joints	0.0660

Figure C1 The performance of pose transfer for six challenging poses, concerning the four metrics compared with the ground truths that were carefully designed by a professional 3D modeler. The highest errors are highlighted in bold.

mask is calculated according to the joints information, represented by 1 channel. During training, we optimize the network by minimizing the cross-entropy loss between the synthesized layout and the ground truth, together with a GAN loss. The training data is obtained from DeepFashion dataset [4] and the layouts are extracted with LIP-JPPNet [5].

Figure D2 shows the difference between layout synthesis results with and without pose transfer. The layout (top) generated by the transferred skeleton of each group yields a shape more similar to the source layout, comparing with the direct result (bottom) without pose transfer.

Appendix E Network Architecture

In Figure E1, we provide more details on the pose encoder, ID encoder, decoder, and discriminator of our pose transfer network, which are modified based on FUNIT [6].

More network configuration information on the pose encoder, ID encoder, decoder, and discriminator of our pose transfer network

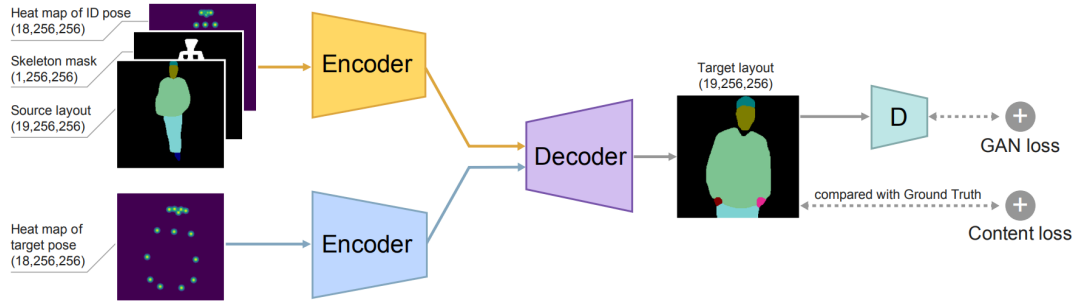


Figure D1 Our network structure for pose-guided layout generation.

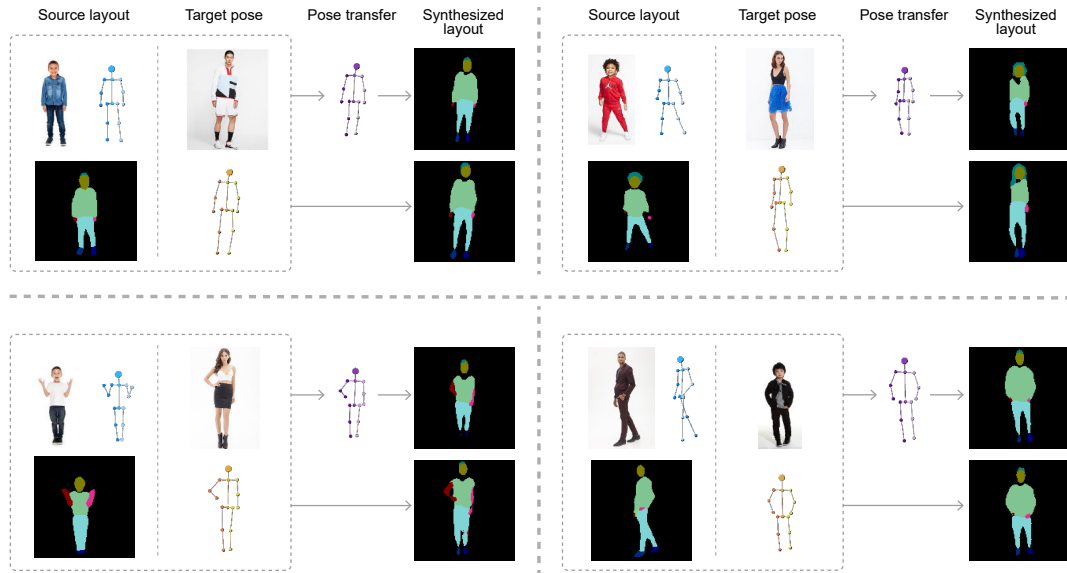


Figure D2 More comparisons for layout synthesis with and without pose transfer.

is illustrated in Table E1.

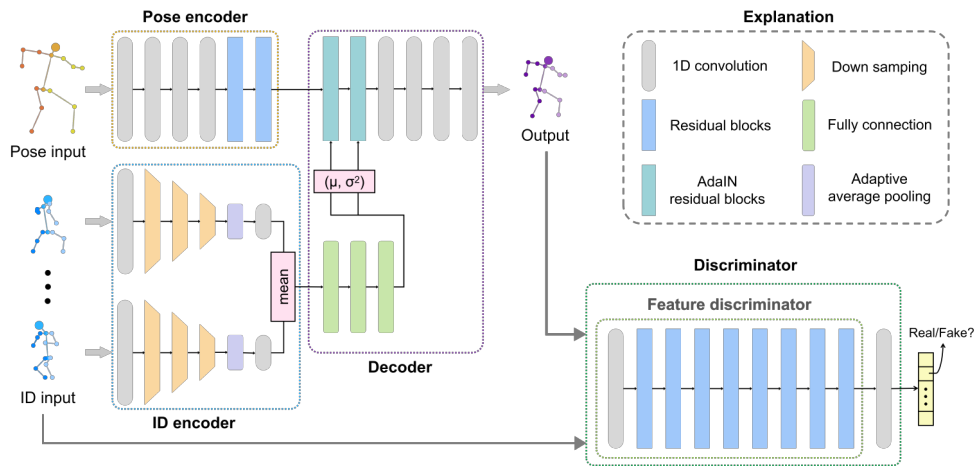


Figure E1 Our detailed pose transfer network architecture.

References

- 1 Adobe Systems Inc. Mixamo. <https://www.mixamo.com>
- 2 Hanbyul Joo, Tomas Simon, Xulong Li, et al. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 2019. 41(1): 190–204

Table E1 Network configuration. Here **K**, **S**, and **P** are the kernel size, stride, and padding. **Channels** denotes the numbers of input and output channels of a convolution layer. **Size** means the size of the input and output feature map. We use one-dimensional convolution layer (Conv1d) in the network. Avgpool, IN, and AdaIN are short for Adaptive Average Pooling, Instance Normalization, and Adaptive IN, respectively. LeakyReLU0.2 represents LeakyReLU activation unit with the slope of 0.2, and N at the bottom row is denoted with the number of ID classes.

	Layer	Details	K	S	P	Channels	Size
Pose Encoder	conv1	Conv1d + IN + ReLU	7	1	3	2/64	15/15
	conv2	Conv1d + IN + ReLU	5	1	2	64/128	15/15
	conv3	Conv1d + IN + ReLU	5	1	2	128/256	15/15
	conv4	Conv1d + IN + ReLU	5	1	2	256/512	15/15
	res1	Conv1d + IN + ReLU	3	1	1	512/512	15/15
		Conv1d + IN	3	1	1	512/512	15/15
	res2	Conv1d + IN + ReLU	3	1	1	512/512	15/15
		Conv1d + IN	3	1	1	512/512	15/15
ID Encoder	conv1	Conv1d + ReLU	7	1	3	2/64	15/15
	conv2	Conv1d + ReLU	4	2	1	64/128	15/7
	conv3	Conv1d + ReLU	4	2	1	128/128	7/3
	conv4	Conv1d + ReLU	4	2	1	128/128	3/1
	avgpool	-	-	-	-	128	1/1
	conv5	Conv1d	1	1	0	128/64	1/1
	Decoder	fc1	FC+ReLU	-	-	-	64/256
fc2		FC+ReLU	-	-	-	256/256	1/1
fc3		FC	-	-	-	256/4096	1/1
res1		Conv1d + AdaIN + ReLU	3	1	1	512/512	15/15
		Conv1d + AdaIN	3	1	1	512/512	15/15
res2		Conv1d + AdaIN + ReLU	3	1	1	512/512	15/15
		Conv1d + AdaIN	3	1	1	512/512	15/15
conv1		Conv1d + IN + ReLU	5	1	2	512/256	15/15
conv2		Conv1d + IN + ReLU	5	1	2	256/128	15/15
conv3		Conv1d + IN + ReLU	5	1	2	128/64	15/15
conv4		Conv1d + Tanh	7	1	3	64/2	15/15
Discriminator		conv1	Conv1d	7	1	3	2/64
	res1	Conv1d + LeakyReLU0.2	3	1	1	64/64	15/15
		Conv1d + LeakyReLU0.2	3	1	1	64/64	15/15
	res2	Conv1d + LeakyReLU0.2	3	1	1	64/64	15/15
		Conv1d + LeakyReLU0.2	3	1	1	64/128	15/15
		Conv1d	1	1	0	64/128	15/15
	res3	Same as res1	3	1	1	128/128	15/15
	res4	Same as res2	3	1	1	128/256	15/15
	res5	Same as res1	3	1	1	256/256	15/15
	res6	Same as res2	3	1	1	256/512	15/15
	res7	Same as res1	3	1	1	512/512	15/15
res8	Same as res2	3	1	1	512/1024	15/15	
conv2	Conv1d + LeakyReLU0.2	1	1	0	1024/ N	15/15	

- Olaf Ronneberger, Philipp Fischer, Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In: Proc. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 2015. 234–241
- Ziwei Liu, Ping Luo, Shi Qiu, et al. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proc. IEEE Conf. on Computer Vision & Pattern Recognition, Las Vegas, NV, USA, 2016. 1096–1104
- Xiaodan Liang, Ke Gong, Xiaohui Shen, et al. Look into person: Joint body parsing & pose estimation network and a new benchmark. IEEE Trans. Pattern Analysis & Machine Intelligence, 2018. 41(4): 871–885
- Ming-Yu Liu, Xun Huang, Arun Mallya, et al. Few-shot unsupervised image-to-image translation. In: Proc. Int. Conf. on Computer Vision, Seoul, Korea, 2019. 10551–10560