Supplementary Material:
Pose-Guided Human Animation from a Single Image in the Wild

Jae Shin Yoon[†] Lingjie Liu[♯] Vladislav Golyanik[♯] Kripasindhu Sarkar[♯]
Hyun Soo Park[†] Christian Theobalt[♯]

[†]University of Minnesota [♯]Max Planck Institute for Informatics, SIC

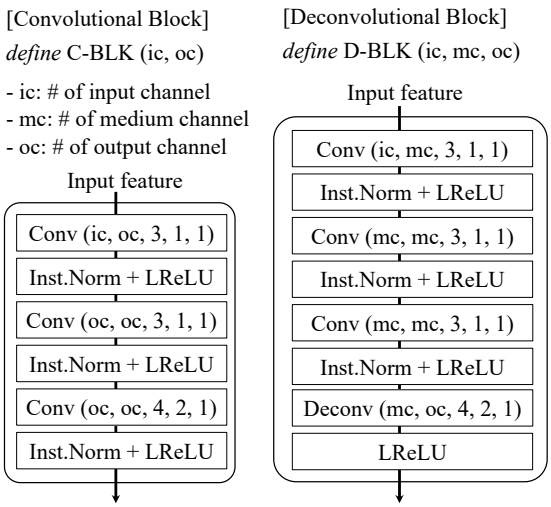

Figure 1. Description of our convolutional and deconvolutional blocks. The convolutional (Conv) and deconvolutional layers (Deconv) take parameters including the number of input channels, the number of output channels, filter size, stride, and the size of zero padding. We use 0.2 for the LeakyReLU (LReLU) coefficient.

This supplementary material provides additional implementation details of our compositional pose transfer network (Sec. A.2), and more results (Sec. B). In the supplementary video, we included the full results of the qualitative comparison, ablation study, more results, and the description of our overall pipeline.

A. Additional Implementation Details

In this section, we provide the implementation details of each modular function in our compositional pose transfer network.

A.1. Network Design

Fig. 2 describes the *SilNet* architecture which takes as input source triplet of the pose map, garment labels, and silhouette, and target pose map, and predicts the silhouette

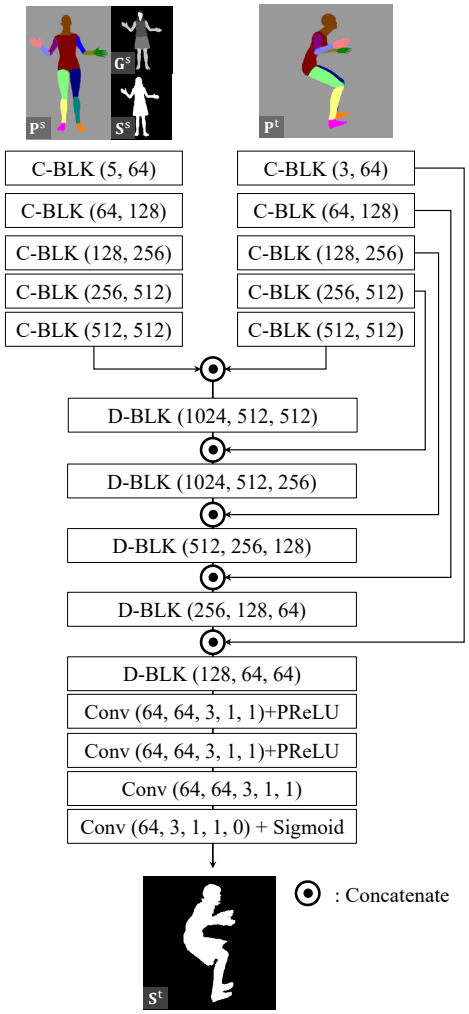

Figure 2. The details of our *SilNet* implementation where C-BLK and D-BLK are described in Fig. 1. Conv and Deconv take as input parameters of (the number of input channels, the number of output channels, filter size, stride, the size of zero padding). We use 0.2 for the LeakyReLU (LReLU) coefficient.

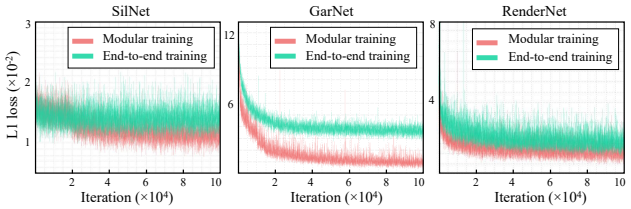

Figure 3. The L1 loss convergence of the modular training (ours) and end-to-end training.

mask in the target pose. Fig. 4 describes the architecture of our *GarNet* that takes as input source triplet of the pose map, silhouette, and garment labels, and target triplet of the pose map, predicted silhouette, and pseudo garment labels, and predicts the complete garment labels. In Fig. 8, we show the details of our *RenderNet* which takes as input source triplet of image, silhouette mask, and garment labels, target silhouette and garment labels, and target pseudo image and its mask, and generates the person image.

For processing each frame in inference time, *SilNet*, *GarNet*, and *RenderNet* take 5ms, 7ms, and 22ms with 1.6GB, 1.6GB, and 2.2GB GPU footprint, respectively, totaling 34ms, or 30 Hz; the memory requirement is 5GB.

A.2. Training Details

We train the proposed *SilNet*, *GarNet*, and *RenderNet* separately in a fully supervised way using only 3D people synthetic dataset [7]. For training, we set the parameters of $\lambda_1 = 0.5$, $\lambda_2 = 0.1$, $\lambda_3 = 0.01$, $\lambda_4 = 10$ and use the Adam optimizer [3] ($lr = 1 \times 10^{-3}$ and $\beta = 0.5$). After training, no further fine-tuning on the testing scene is required.

Our networks are differentiable, and thus, end-to-end trainable. However, since each task is heavily disjointed (e.g., using different ground-truth labels), the end-to-end training does not result in meaningful improvement. Further, we empirically found that due to the dependency between networks, the loss convergence of the end-to-end training is suboptimal as shown in Fig. 3. In fact, the modular training achieves lower error for each task. We will include a convergence analysis in the revised version.

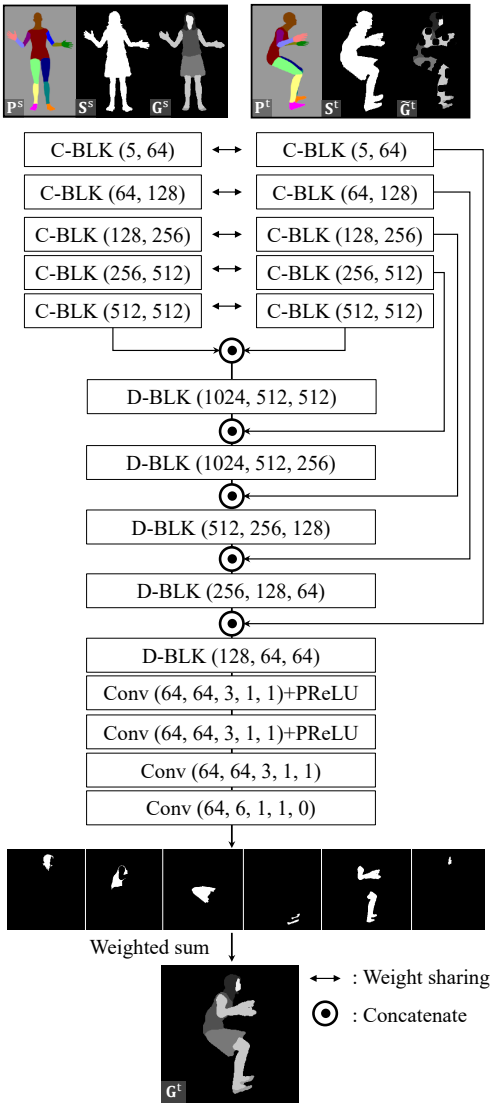

Figure 4. The details of our *GarNet* implementation where C-BLK and D-BLK are described in Fig. 1. Conv and Deconv take as input parameters of (the number of input channels, the number of output channels, filter size, stride, the size of zero padding). We use 0.2 for the LeakyReLU (LReLU) coefficient.

B. More Results

B.1. Additional Dataset Description

We provide more details of the videos used for the evaluation. In order to evaluate our approach, we use eight sequences of the subjects in various clothing and motions from existing works [9, 8, 4, 1, 2] and capture two more sequences which include a person with more complex clothing style and movement than others. *RoM1* and *RoM2*: Two men show their range of motion with various poses [8]. *Jumping* [9]: A woman in a black and white coat jump from one side to another. *Kicking* and *Onepiece* [2]: A man and woman take the motion of kicking and dancing where the woman is wearing a unique onepiece. *Checker* [4]: A man in shirts with checkered pattern swings his hands. *Rotation1* and *Rotation2* [1]: Two A-posed men rotate their body. *Maskman*: A man wearing a facial mask shows his various motion. *Rainbow*: A woman in a sweater with rainbow pattern turns her body with dancing motion.

B.2. User Study Results

In our user study, three questions are asked: Q1: Which video looks most realistic including temporal coherence? Q2: Which video preserves the identity best including fa-

cial details, shape, and overall appearance? Q3: In which video, the background is preserved better across the frames (only for the case of scenes with background)? For each method, we measure the performance based on the number of entire votes divided by the number of participants and the number of occurrence in the questionnaires. The full results are shown in Fig. 5. The first question was answered in 84.3% and 93.0% of the cases in favour of our method with and without the ground truth sequence, respectively, and the second question 84.1% and 94.2%. In the third question, the background is preserved better in our method than LWG in 96.8 % of the answers. The results show that our method outperforms other state of the art, and our animations are in many cases qualitatively comparable to real videos of the subjects. The choice between a real video and our animation did not fall easy because the ground-truth video often contains noisy boundary originated from the person segmentation error while the generated person images from our method shows the clear boundary.

B.3. Additional Quantitative Results

We include the quantitative results which do not appear in the main paper. In Table 1, the performance of the baseline models that are pretrained from the DeepFashion (DF) dataset by the authors is summarized in the first chunk (from 2 to 6 row), ablation study in the second chunk (from 7 to 16), and application to the multiview data in the third chunk (from 17 to 18).

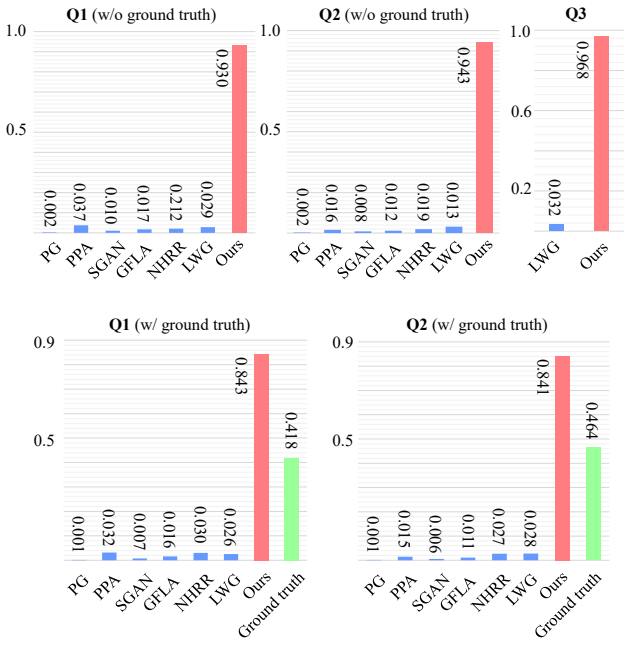

Figure 5. The full results of the user study where *x*-axis represents the number of votes for the associated method which is normalized by the number of participants and the number of occurrence in the questionnaires. Q1, Q2, and Q3 represent the question type. Our results were often ranked as more realistic than the real videos because they involve a significant boundary noise from the person segmentation error while our method produces the human animation with clean boundary.

References

[1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *CVPR*, 2018. 2

[2] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *CVPR*, 2020. 2

[3] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv e-prints*, 2014. 2

[4] 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 *ICCV*, 2019. 2

[5] Arun Mallya, Ting-Chun Wang, Karan Sapra, and Ming-Yu Liu. World-consistent video-to-video synthesis. *ECCV*, 2020. 4

[6] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 4

[7] Albert Pumarola, Jordi Sanchez, Gary Choi, Alberto Sanfeliu, and Francesc Moreno-Noguer. 3DPeople: Modeling the Geometry of Dressed Humans. In *ICCV*, 2019. 2

[8] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *SIGGRAPH Asia*, 2020. 2

[9] Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz. Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In *CVPR*, 2020. 2

	Maskman	Rainbow	RoM1	RoM2	Jumping	Kicking	Onepiece	Checker	Rotation1	Rotation2	Average
PG (DF)	2.01 / 4.24	2.14 / 4.41	2.22 / 4.48	1.81 / 4.31	2.33 / 4.32	2.15 / 4.49	2.43 4.66	2.07 / 4.25	1.74 / 4.18	2.58 / 4.47	2.15 / 4.38
SGAN (DF)	2.33 / 3.96	2.39 / 4.22	2.50 / 4.16	2.12 / 4.22	2.63 / 4.09	2.49 / 4.29	2.67 / 4.25	2.34 / 3.99	1.89 / 3.93	2.74 / 4.22	2.43 / 4.13
PPA (DF)	2.84 / 3.76	2.70 / 3.80	2.78 / 3.91	2.65 / 3.97	2.89 / 3.87	2.88 / 3.94	3.21 / 4.05	2.26 / 3.76	2.26 / 3.75	3.01 / 3.77	2.74 / 3.86
GFLA (DF)	1.96 / 3.86	1.64 / 3.93	2.19 / 3.89	1.50 / 3.99	2.01 / 3.85	2.05 / 3.96	2.23 / 3.94	1.74 / 3.84	1.60 / 3.88	1.92 / 3.89	1.88 / 3.90
NHHR (DF)	1.71 / 2.96	1.89 / 3.06	1.82 / 3.07	1.56 / 3.03	2.06 / 3.03	1.68 / 3.11	2.16 / 3.16	1.48 / 2.94	1.80 / 3.02	2.77 / 3.11	1.89 / 3.05
R	1.64 / 2.31	1.48 / 2.43	1.41 / 2.30	1.53 / 2.44	2.00 / 2.54	1.16 / 2.18	1.36 / 2.34	1.41 / 2.31	1.22 / 2.31	1.62 / 2.33	1.48 / 2.35
GR	1.64 / 2.30	1.45 / 2.42	1.51 / 2.30	1.44 / 2.42	1.91 / 2.53	1.40 / 2.24	1.24 / 2.35	1.39 / 2.29	1.21 / 2.30	1.60 / 2.32	1.47 / 2.35
SR	1.57 / 2.26	1.30 / 2.42	1.31 / 2.24	1.41 / 2.37	1.89 / 2.54	1.17 / 2.20	1.24 / 2.33	1.11 / 2.24	1.05 / 2.23	1.25 / 2.22	1.33 / 2.31
SGR-$\mathbf{S}^{\mathrm{s}}$	1.58 / 2.29	1.33 / 2.41	1.26 / 2.26	1.43 / 2.39	1.99 / 2.54	1.18 / 2.23	1.29 / 2.35	1.10 / 2.36	1.05 / 2.23	1.24 / 2.20	1.35 / 2.32
SGR-$\mathbf{G}^{\mathrm{s}}$	1.66 / 2.30	1.38 / 2.39	1.31 / 2.32	1.48 / 2.35	1.89 / 2.51	1.18 / 2.23	1.31 / 2.40	1.31 / 2.31	1.19 / 2.28	1.42 / 2.30	1.41 / 2.34
SGR-$\bar{\mathbf{I}}^{\mathrm{t}}$	1.79 / 2.28	1.97 / 2.49	1.55 / 2.30	1.52 / 2.38	2.13 / 2.50	1.31 / 2.23	1.79 / 2.39	1.49 / 2.31	1.15 / 2.22	1.50 / 2.21	1.62 / 2.33
SGR-$\mathbf{z}^{\mathrm{s}}$	1.57 / 2.27	1.31 / 2.40	1.25 / 2.26	1.42 / 2.38	1.90 / 2.52	1.15 / 2.19	1.29 / 2.31	1.11 / 2.22	1.05 / 2.19	1.24 / 2.23	1.32 / 2.30
SGR-L_{KL}	1.54 / 2.27	1.25 / 2.38	1.27 / 2.25	1.40 / 2.38	1.88 / 2.55	1.13 / 2.19	1.25 / 2.32	1.09 / 2.24	1.04 / 2.20	1.15 / 2.19	1.30 / 2.30
SGR-$\mathbf{A}$	1.59 / 2.28	1.28 / 2.40	1.31 / 2.26	1.40 / 2.38	1.86 / 2.51	1.23 / 2.21	1.32 / 2.33	1.14 / 2.25	1.15 / 2.23	1.28 / 2.20	1.36 / 2.31
SGR (**full**)	**1.54 / 2.27**	**1.24 / 2.38**	**1.25 / 2.24**	**1.38 / 2.36**	**1.87 / 2.53**	**1.08 / 2.19**	**1.23 / 2.32**	**1.09 / 2.24**	**1.00 / 2.19**	**1.12 / 2.16**	**1.28 / 2.29**
SGR+2view	1.50 / 2.25	1.22 / 2.38	1.21 / 2.23	1.33 / 2.36	1.80 / 2.51	1.15 / 2.17	1.20 / 2.31	1.07 / 2.23	0.97 / 2.16	1.06 / 2.14	1.25 / 2.28
SGR+4view	1.49 / 2.25	1.21 / 2.38	1.21 / 2.23	1.33 / 2.35	1.80 / 2.51	1.12 / 2.17	1.20 / 2.31	1.07 / 2.23	0.98 / 2.16	1.07 / 2.14	1.24 / 2.27

Table 1. Quantitative results with LPIPS (left, scale: 10^{-1}) and CS where the lower is the better. We denote the full model used for the comparison with other baseline methods as SGR (**full**).

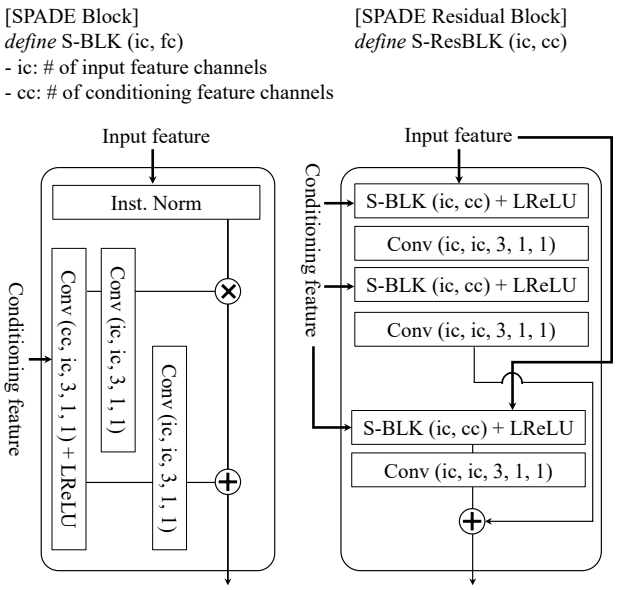

Figure 6. The description of SPADE and SPADE Residual blocks similar to [6]. Conv take as input parameters of (the number of input channels, the number of output channels, filter size, stride, the size of zero padding). We use 0.2 for the LeakyReLU (LReLU) coefficient.

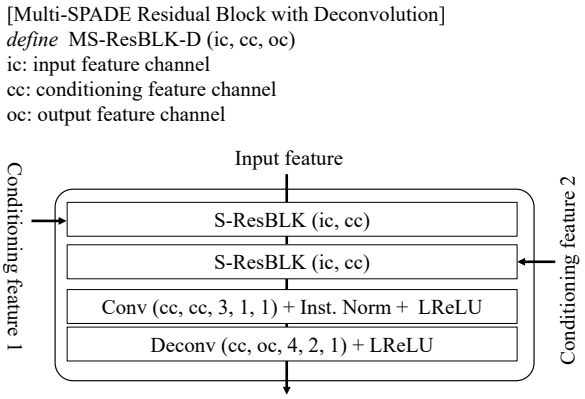

Figure 7. The description of Multi-Spade blocks similar to [5] where the details of S-ResBLK is described in Fig. 6. Conv and Deconv take as input parameters of (the number of input channels, the number of output channels, filter size, stride, the size of zero padding). We use 0.2 for the LeakyReLU (LReLU) coefficient.

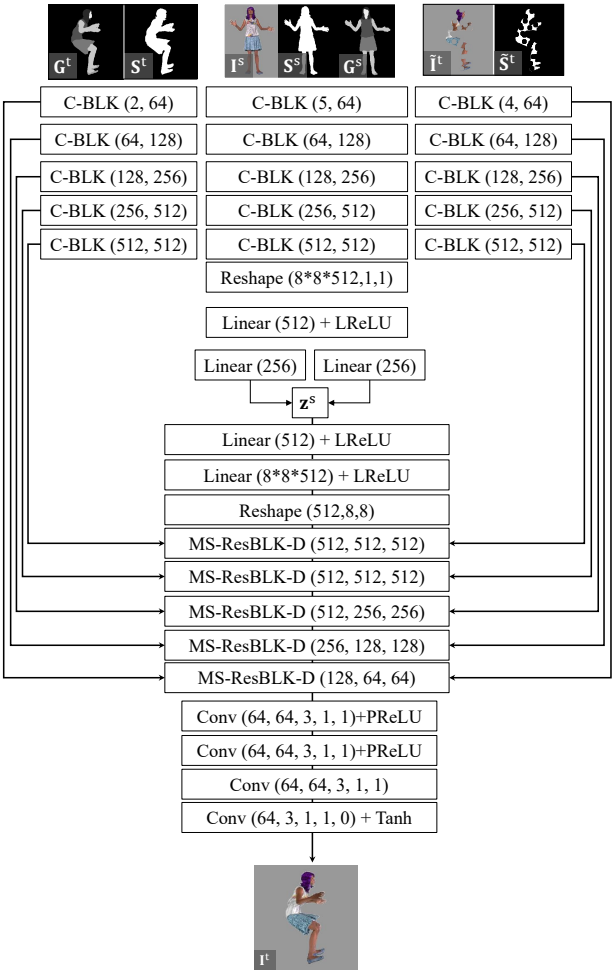

Figure 8. The details of our *RenderNet* where C-BLK and D-BLK are described in Fig 1, and MS-ResBLK-D is in Fig. 7. Conv takes as input parameters of (the number of input channels, the number of output channels, filter size, stride, the size of zero padding). We use 0.2 for the LeakyReLU (LReLU) coefficient.