High-Fidelity Neural Human Motion Transfer from Monocular Video — Supplementary Material —

Moritz Kappel[1] Vladislav Golyanik[2] Mohamed Elgharib[2] Jann-Ole Henningson[1]
Hans-Peter Seidel[2] Susana Castillo[1] Christian Theobalt[2] Marcus Magnor[1]

[1] Computer Graphics Lab, TU Braunschweig, Germany {lastName}@graphics.tu-bs.de
[2] Max Planck Institute for Informatics, Saarland Informatics Campus, Germany {lastName}@mpi-inf.mpg.de

In this supplementary document, we provide more details on the methods examined in the main paper along with additional results for the presented framework.

1. Garment Conditioning Representation and Visualization

Our current implementation uses the self-correction for human parsing method by Li *et al*. [7], that was trained on the ATR dataset [8]. Thus, our method currently supports a total of eighteen different labels (namely background, hat, hair, sunglasses, upper clothes, skirt, pants, dress, belt, left shoe, right shoe, face, left leg, right leg, left arm, right arm, bag and scarf). To generate training data and visualize the results, we use the official code of [7] provided by the authors[1]. Naturally, our framework is compatible with arbitrary human parsing methods, and can easily be modified to support the latest state-of-the-art approaches as well as new datasets including labels for different types of clothing.

To estimate the internal gradient structure, we adopt the procedure described by Tan *et al*. [12] for conditioning hair structure. More precisely, we extract ground-truth annotations from the video sequence using a set of 32 oriented Gabor-filters K_Θ with $\Theta = [0, \pi)$ being the discrete angle values. By applying the filter stack to every pixel, we extract a dense orientation o_n and confidence c_n map for image i_n by calculating the angle and amplitude of the maximum filter response as:

$$\begin{aligned} o_n &= \arg\max_{\Theta} |\,(K_\Theta \otimes i_n)\,|, \\ c_n &= \max_{\Theta} |\,(K_\Theta \otimes i_n)\,|, \end{aligned} \tag{1}$$

where $\otimes$ denotes the convolution operator. We then follow their procedure of converting o_n to a continuous representation, and Gaussian-filter the orientation map based on the local confidence c_n to reduce noise. However, in contrast to hair, which comprises a dense structure field by nature, the local gradient structure of clothing is usually sparse (*e.g*., tight monochromatic cloth) and highly dependent on the texture and material. Thus, unlike the original implementation that discards confidence after filtering, we normalize c_n in the range $[0, 1]$, and append it to the orientation map, which results in our final two-channel clothing structure representation $w_n = (o_n, c_n)$. Thus, intuitively, the first channel of our clothing structure tells the appearance network in which direction a gradient (*e.g*., produced by a wrinkle) points, while the second channel models the confidence and strength of texture changes for each individual pixel. In our figures and supplemental video, we visualize the structure representation w_n similar to optical flow vectors, where the orientation and length (confidence) are modeled as hue and saturation in HSV-color space, respectively, as shown in Fig. 1.

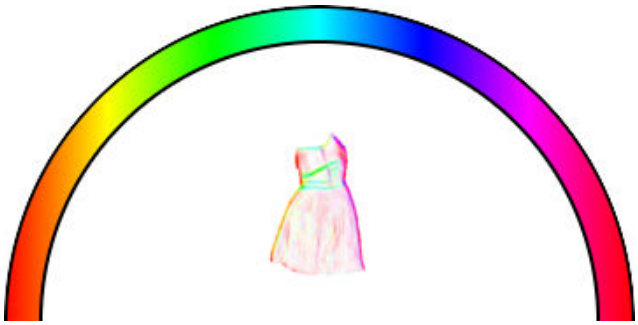

Figure 1. Example of our garment structure visualization for one of the dresses shown in the teaser from the main paper. We model the orientation as hue in HSV-color space, while each pixel's saturation corresponds to the local gradient confidence.

2. Perceptual Experiment

In our perceptual experiments, we aim at comparing the reenacting results from an observer's perspective, which requires multiple stimuli with differences between them often

[1] https://github.com/PeikeLi/Self-Correction-Human-Parsing

being quite subtle. More importantly, the quality of the results that we aim to measure cannot be represented on a linear scale [5], which advises against ranking the methods. Therefore, we chose the paired comparisons technique, where the participants are shown two reenacted videos at a time, side by side, and are asked to choose the one that better fits the task question. We thus performed a two-alternatives-forced-choice (2AFC) preference task assessing reenactments by two compared methods for a given video.

2.1. Stimuli

The First Experiment – Our Dataset. As described in the main paper, we trained our method along with other three state-of-the-art reenacting techniques (EDN [3], pix2pixHD [13], and Recycle-GAN [1]) on some sequences of our data set (see the top part of Fig. 2).

The Second Experiment – Liu *et al.*'s Dataset. We also evaluated our technique on the dataset of Liu *et al.* [9] (see the second bottom part of Fig. 2). Here, we trained two of the compared methods (EDN [3] and ours) on this dataset. Furthermore, we used the results of the reenactment approach of Liu *et al.* that was provided by the authors. Note we can not replicate their method on any other dataset as it requires a sophisticated capturing pipeline using a monocular structure-from-motion reconstruction of the target actor, including delicate pre-processing and mesh fitting. For fair comparison, all results were down-sampled to the resolution given by Liu *et al.* (256x256). Given the set of videos and the three tested methods (Liu [9], EDN [3] and ours), the total number of possible paired comparisons reduces to twelve, making it well suitable for a single participant to perform a complete test while maintaining the necessary level of attention.

2.2. Experimental Procedure

The First Experiment. The study was hosted online, and a total of 54 subjects from various computer science backgrounds participated in the study. Before starting the experiment, participants were presented a short text describing the task and the procedure. The subjects were exposed to several video pairs that play side by side. Each video in a pair was produced by a different technique, and the order of pairing between methods, position on the screen and order of display of the pairs was fully randomized. For each video pair, the subjects were asked to record their answers to two questions: Q1 (*"Which video looks more realistic?"*) and Q2 (*"Which video shows more natural motion and deformation of the clothes?"*). The study included a total of twelve video pairs and took around ten minutes to complete.

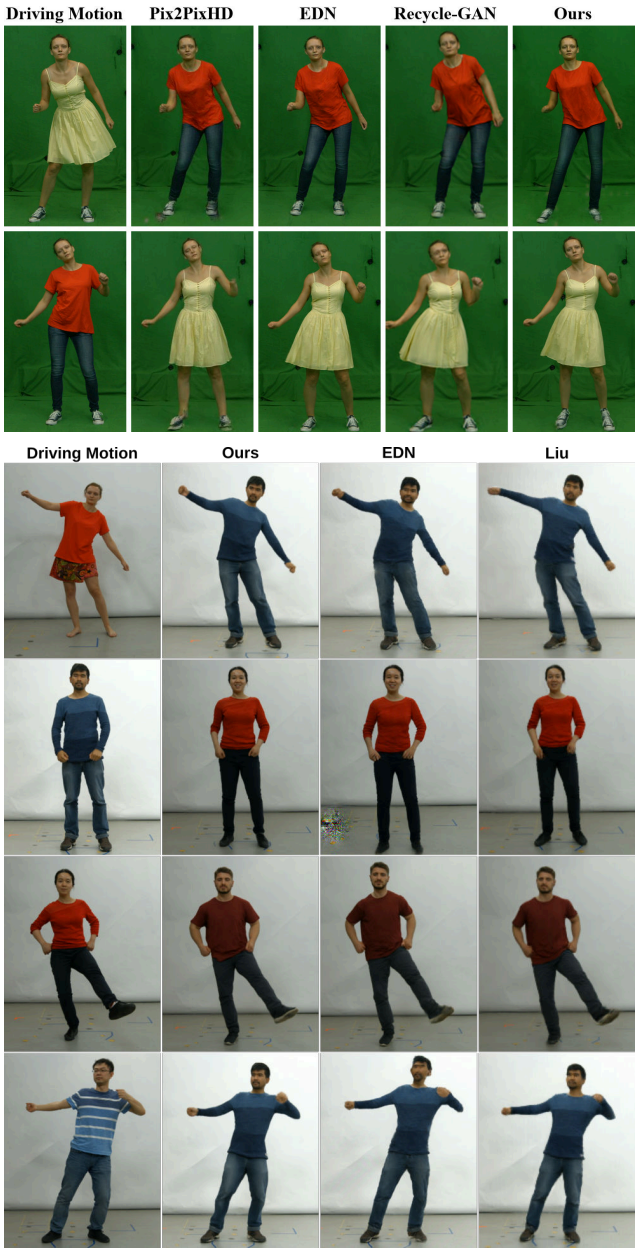

Figure 2. Exemplary frames from the video sequences used as stimuli for our perceptual experiments. Top: *Exp. 1* conducted on our dataset. Bottom: *Exp. 2* on the dataset of Liu *et al.* [9].

The Second Experiment. The second experiment was conducted *in-situ*. A total of 16 people (graphics and vision experts) participated (age range 22-40 years; four women). The mean time to complete the experiment was nine minutes. Before the experiment began, each participant was informed about the structure and flow of the experiment—but not of the research question behind it—and was given the option to ask any further questions. The participants performed the experiment one at a time. Before the exper-

imenter left the room, the participant was asked to sit in a semi-dark room roughly 50 cm in front of a 24" LED monitor (with the resolution of 1920×1080). The respondents then were exposed to a screen describing in detail the instructions and were given another chance to ask questions. The experiment was controlled by Psychophysics toolbox, version 3.0.15 (PTB-3) [2, 11, 6]. At the start of each trial, participants were presented with two videos side by side. Participants were explicitly instructed to focus their attention on the garments and ignore any potential artifact not concerning the clothes. The participants were able to enter their answers by clicking on the desired video and also to replay each video individually as many times as desired; once a response was entered, the next trial started. Both the order of the stimuli, their pairing and their position on the screen (left *vs* right) were fully randomized, with each participant receiving a different random order. Each participant reported normal or corrected-to-normal vision.

2.3. Analysis

To measure not only the performance of each method but the agreement between participants, we followed the linked-paired comparison design [4]. Thus, we rank methods according to the number of times they are preferred over the other methods. The total number of votes a method received is displayed in Table 1 for the first experiment ("*Exp. 1*") and Table 2 for the second experiment ("*Exp. 2*").

To analyze the true meaning of this ranking, we perform a significance test of the score differences. Towards that goal, we need to find a value R' for which the variance-normalized range of scores within each group is lower or equal to that value. This means that we need to compute R' such that $P[R \geq R'] \leq \alpha$, where α is the confidence level, which we set to 0.01. Then, following the work of David [4] we can derive R' from

$$P\left(W_{t,\alpha} \geq \frac{2R' - 0.5}{\sqrt{mt}}\right), \tag{2}$$

where t is the number of methods to be compared, m is the number of participants and $W_{t,\alpha}$ has been previously tabulated by Pearson and Hartley [10]. In our case, $W_{4,0.01} = 4.405$ for *Exp. 1* (Table 1) and $W_{3,0.01} = 4.125$ for *Exp. 2* (Table 2). This leads us to the values $R'_{Exp1} = 32.62001$ for the first experiment and $R'_{Exp2} = 14.53942$ for the second experiment. Since all the differences between the ranked groups are bigger than the obtained R', we can conclude that they are all statistically significant. Thus, the ranking creates four distinguishable groups in *Exp. 1* for both Q1 (*"Which video looks more realistic?"*) and Q2 (*"Which video shows more natural motion and deformation of the clothes?"*). Furthermore, the ranking creates three distinguishable positions in *Exp. 2* (*"Which video displays more realistic clothes? (movement/deformations/appearance)"*).

Exp. 1	Method	Input	#Votes	Ranking
Q1	**Ours**	Video (Pose)	257	**1**
	Recycle-GAN [1]	Video (RGB)	194	**2**
	EDN [3]	Video (Pose)	143	**3**
	pix2pixHD [13]	Video (Pose)	54	**4**
Q2	**Ours**	Video (Pose)	243	**1**
	Recycle-GAN [1]	Video (RGB)	205	**2**
	EDN [3]	Video (Pose)	135	**3**
	pix2pixHD [13]	Video (Pose)	65	**4**

Table 1. *Exp. 1*: Perceptual ranking of the compared methods for our online experiment with 54 participants with various backgrounds for each of the questions (Q1 and Q2). All the rankings are statistically significant.

Exp. 2	Method	Input	#Votes	Ranking
	Liu [9]	Textured mesh + Video (Pose)	103	**1**
	Ours	Video (Pose)	76	**2**
	EDN [3]	Video (Pose)	13	**3**

Table 2. *Exp. 2*: Perceptual ranking of the compared methods on the in-situ user-study with 16 CG/CV experts. The rankings are statistically significant. This table is already included in the main paper at the bottom of Table 2 and is repeated here for convenience.

Video Attributes. Additionally, to gain more insight on the reasons to choose one result over another, in the experiment conducted *in-situ* with CG/CV experts, the participants were occasionally asked to pick one or several items out of a proposed set of reasons for not choosing a result. This question appeared randomly with the probability of $1/3$, *i.e.,* the frequency we found suitable in order to maintain the participant's attention without making the test tedious. Table 3 shows the complete list of reasons and how often they were selected as a reason for rejecting a given result. As can be derived from the table, the most frequent reasons to discard a method were artifacts and implausible deformations of the clothing.

This is the video you did not choose in the last comparison. Please specify which of the following bothers you in this video. You may check multiple options.	% cases reason selected
Unrealistic wrinkles.	37.5%
Unrealistic clothes' texture.	23.44%
Implausible deformations of the clothing.	**59.63%**
Temporal inconsistencies in the clothes' movement.	20.31%
Other artifacts in clothing.	**45.31%**
The other result was simply more appealing.	10.94%
Other.	3.21%

Table 3. Questionnaire displayed after rejecting a result. The participant was able to choose as many answers as desired. The second column shows the frequency (%) for reporting one type of artifact when the questionnaire appeared after rejecting a result (four times per participant). These numbers indicate the most frequent reasons to discard a method, and thus, the most important features for participants.

3. Training Details

Our framework $\boldsymbol{G}_\theta$ consist of four trainable components, namely $\boldsymbol{G_{ref}}$, $\boldsymbol{G_{app}}$, $\boldsymbol{G_{str}}$, $\boldsymbol{G_{shp}}$. For every network, we adapt the local Pix2PixHD [13] generator architecture, resulting in memory consumption and training times of $\sim 3.5\times$ the reported values for the original implementation ($\boldsymbol{G_{ref}}$ uses a reduced amount of intermediate blocks due to the simplicity of the learned mapping). However, to enable training on current graphics hardware, every instance is currently trained individually using annotations and losses described in the main document, which allows for stepwise processing on consumer-grade devices, requiring $\sim$5GB VRAM and $\sim$280 milliseconds per frame. To achieve temporal smoothness in our garment conditioning modules—and, thus, temporally consistent renderings of body parts and clothing—we condition the shape and structure predictions on the previous outputs of the networks. Intuitively, this information is crucial to estimate the non-rigid deformations within clothing, as they do not only depend on forces extracted from the change of pose, but also the current state of the dynamic system. Thus, instead of computing two consecutive frames and applying a temporal discriminator, we process the videos in an entirely sequential manner, which results in temporally more stable results, as shown in our supplemental video. Still, we cut the gradient to previous outputs, as we find that truncated temporal backpropagation significantly increases memory consumption and training times, without contributing much to the final quality.

To further stabilize initial training or our recurrent components, we use a combination of teacher forcing and conventional training, similar to curriculum learning. More specifically, we use (pseudo) ground-truth annotations for the first epoch to stabilize and speed up initial training, which encourages the network to rely upon accurate estimates of the last time step. During later epochs, we feed back previous network outputs to relax the reliance on faultless input data, helping the network to generalize to new motion patterns and recover from erroneous predictions. We further synthesize the first frame, where no predecessor is available, based on a black image and iteratively execute our networks on the initial input pose (with temporal derivatives set to zero) until the output converges towards a reasonable estimate (in practice, this takes around 20-30 iterations).

References

[1] Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. Recycle-gan: Unsupervised video retargeting. In *European Conference on Computer Vision (ECCV)*, pages 119–135, 2018. 2, 3

[2] David H. Brainard. The psychophysics toolbox. *Spatial Vision*, 10:433–436, 1997. 3

[3] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *International Conference on Computer Vision (ICCV)*, 2019. 2, 3

[4] Herbert Aron David. *The method of paired comparisons*, volume 12 of *Griffin's statistical monographs & courses*. Charles Griffin, London, 1 1963. 3

[5] Maurice G. Kendall and Bernard Babington Smith. On the method of paired comparisons. *Biometrika*, 31(3-4):324–345, 03 1940. 2

[6] Mario Kleiner, David Brainard, and Denis Pelli. What's new in psychtoolbox-3? *Perception 36, European Conference on Visual Perception (ECVP) Abstract Supplement*, 2007. 3

[7] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *arXiv preprint arXiv:1910.09777*, 2019. 1

[8] Xiaodan Liang, Si Liu, Xiaohui Shen, Jianchao Yang, Luoqi Liu, Jian Dong, Liang Lin, and Shuicheng Yan. Deep human parsing with active template regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(12):2402–2414, 2015. 1

[9] 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 (TVCG)*, 2020. 2, 3

[10] Egon S. Pearson and Herman O. Hartley. *Biometrika Tables for Statisticians*, volume 1. Cambridge University Press, 3 edition, 1966. 3

[11] Denis G. Pelli. The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, 10:437–442, 1997. 3

[12] Zhentao Tan, Menglei Chai, Dongdong Chen, Jing Liao, Qi Chu, Lu Yuan, Sergey Tulyakov, and Nenghai Yu. Michigan: multi-input-conditioned hair image generation for portrait editing. *ACM Transactions on Graphics (TOG)*, 39(4):95–1, 2020. 1

[13] 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 *Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3, 4