High-Fidelity Neural Human Motion Transfer from Monocular Video

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]

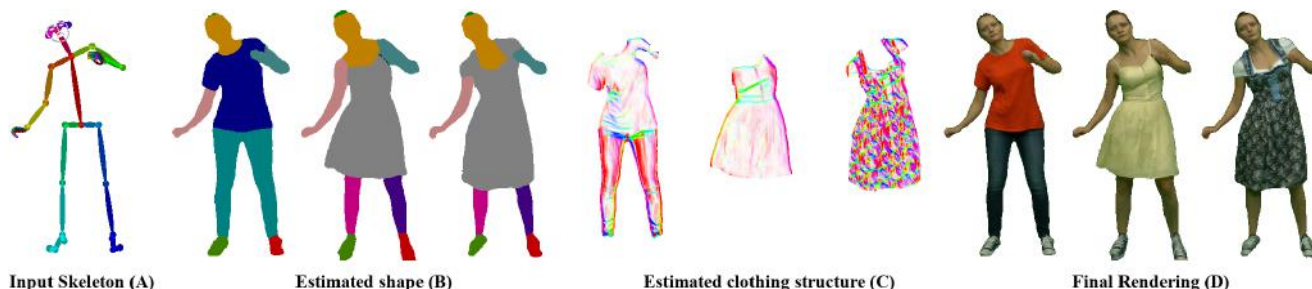[1]ICG, TU Braunschweig   [2]MPI for Informatics, SIC

Figure 1. **Our human motion transfer framework** translates 2D pose skeletons (**A**) to photorealistic images (**D**) by explicitly estimating the actor's 2D shape (**B**) and internal structure of clothing (**C**), enabling direct deformation handling and support of high-frequency details such as wrinkles and shadows.

Abstract

Video-based human motion transfer creates video animations of humans following a source motion. Current methods show remarkable results for tightly-clad subjects. However, the lack of temporally consistent handling of plausible clothing dynamics, including fine and high-frequency details, significantly limits the attainable visual quality. We address these limitations for the first time in the literature and present a new framework which performs high-fidelity and temporally-consistent human motion transfer with natural pose-dependent non-rigid deformations, for several types of loose garments. In contrast to the previous techniques, we perform image generation in three subsequent stages, synthesizing human shape, structure, and appearance. Given a monocular RGB video of an actor, we train a stack of recurrent deep neural networks that generate these intermediate representations from 2D poses and their temporal derivatives. Splitting the difficult motion transfer problem into subtasks that are aware of the temporal motion context helps us to synthesize results with plausible dynamics and pose-dependent detail. It also allows artistic control of results by manipulation of individual framework stages. In the experimental results, we significantly outperform the state-of-the-art in terms of video realism. Our code and data will be made publicly available[1].

[1]https://graphics.tu-bs.de/publications/
kappel2020high-fidelity

1. Introduction

Human motion transfer methods, also known as performance cloning or reenactment methods, can generate realistic video animations of an actor following a target motion specified by a user. This has several applications in AR/VR and video editing. Building upon new advances in machine learning, current motion transfer methods tackle this challenging problem by learning a direct mapping between an actor-independent motion space and the resulting target actor's appearance space. These methods often require a training video of an actor performing a rich set of motions [6, 32, 22, 20, 2, 38].

Some recent motion transfer approaches parameterize motion as skeletal pose sequences that can be computed from videos with off-the-shelf pose detectors [6, 32]. Others use pre-captured template meshes or parameterized body models to provide additional guidance to the synthesis step [22, 20]. Acquisition of such templates [20], however, requires an extensive structure-from-motion reconstruction of the static target actor under constant lighting. Furthermore, existing human motion transfer approaches are likely to produce notable temporal and spatial artifacts when actors wear loose clothing, such as dresses, skirts and hoodies [6, 2, 38, 22]. On such garments, they struggle to realistically reproduce the appearance of fine-scale details like folds and wrinkles, as well as plausible dynamics.

In this paper, we present a new human motion transfer

framework that generates visually plausible video animations of humans that are spatially and temporally coherent, and show natural dynamics, even for actors wearing loose garments (see Fig. 1). Given a single monocular video of an actor performing a rich set of motions, we train a stack of deep generative networks to learn a mapping from 2D pose to a silhouette with semantic part labels, and per-pixel appearance of the actor. We model the person's shape as a dense foreground silhouette mask with per-pixel labels encoding assignment to limbs and garments. We further encode the structure of wrinkles and texture patterns of garments as the orientation and strength of local image gradients. We extract this structure from images using a bank of oriented filter kernels [26, 36]. Encoding the actor's appearance with these explicitly decoupled intermediate representations of silhouette and structure is key to enhance the temporal and spatial quality of synthesized videos comprising human actors in loose clothing.

Our method improves over current motion transfer approaches in terms of visual fidelity using a single RGB camera. Furthermore, our representation provides an additional level of control over the final image generation. For example, for the same overall dynamic geometric outline (*i.e.,* the same garment geometry), color and appearance, including fold and wrinkle style, can be manipulated in a purely image-based way. Overall, our contributions can be summarized as follows: (1) A new motion transfer framework with an emphasis on visually-plausible fine-scale deformations and dynamics in the actor's clothing. (2) For this, we propose to decompose the pose-to-image translation task into better conditioned cascaded processes, where the final appearance is conditioned on the predicted shape outline and internal structure of the clothing. (3) We show that our intermediate representations do not only help to provide more temporally coherent conditioning resulting in more appealing image synthesis, but also allow controlling individual aspects of the final rendering (*e.g.,* enhance wrinkles and transfer the clothing style).

2. Related Work

Apart from general-purpose image-translation techniques which can be used for novel human view synthesis [13, 42, 2, 38, 37], most specialized techniques can be classified into still image-based [30, 33, 24, 25, 9] or video-based [6, 1, 8, 21, 20, 22, 10]. Pix2pix [13] is one of the most recognized general-purpose image translation techniques in the literature. Other image-to-image translation variants include pix2pixHD [38] for processing high-resolution images, Zhu *et al.* [42] for learning from unpaired data, and Wang *et al.* [37] for video processing. The work of Bansal *et al.* [2] learns a mapping from unpaired video data with a focus on processing temporal information. While general-purpose image-to-image translation

techniques [13, 42, 2, 38, 37] can in principle be applied in our scenario, their visual accuracy usually suffers from multiple types of artifacts such as spurious details (*e.g.,* jittering texture details), unnatural deformations, and temporal inconsistency, among others.

Neural human motion transfer methods are specifically designed to generate novel views of a person observed in a single RGB image. Still-image-based approaches [30, 33, 24, 25, 9, 8] focus is different from the problem we solve. Their focus is on re-targeting a person in just a still image and do not produce temporally consistent and visually pleasing video results. Video-based human motion transfer techniques [6, 1, 21, 20, 22, 10], however, focus on producing video results by handling temporal information. Aberman *et al.* [1] combine a skeleton representation with an image-translation network for human performance cloning. Additionally, they introduce an actor-independent training branch for unpaired data to enlarge the set of representable poses. Recently, Chan *et al.* [6] introduced a translation method for human bodies using an intermediate, subject independent pose representation. Inspired by pix2pix [13], they learn a pose-to-appearance mapping from video clips of target identities for human performance transfer. To improve temporal consistency and overall image quality, they include a temporal discriminator and dedicated face prediction network. While their method produces visually appealing results, it frequently generates unnatural and inconsistent deformations and texture details for loose clothing, as it is underconstrained from 2D pose keypoints.

Some video-based techniques aim to improve temporal and spatial fidelity by incorporating 3D information through human body meshes [21, 20, 22, 10]. L. Liu *et al.* [21, 20] presented high-quality human reenactments. As a proxy, they use an explicitly rigged, textured and skinned 3D model of the target actor. The underlying deformation model, however, does not allow to capture small wrinkles and local deformations. W. Liu *et al.* [22] use a human mesh recovery technique to disentangle poses from shapes as well as a correspondence map for human appearance transfer and novel view synthesis. They propagate input information both in the image and feature space and advocate a warping-based module for the enhanced preservation of the source information. Along the same lines is the work of Guan *et al.* [10] for human action transfer. It shows generalization to different target persons without retraining, thanks to a texture extraction method and a parametric human body model. However, the resolution of the supported renderings does not capture fine details and wrinkles. Textured Neural Avatars [32] can be trained for rendering a target actor in tight clothes in arbitrary body and camera poses. Proceeding from monocular videos with extracted 3D poses and regions of interest (human bodies), they train a convolutional neural network to predict a dense UV-map under a

given target pose. Explicit decoupling of the geometry and texture results in higher visual quality and temporal consistency; yet, the overall representation is still comparably coarse and lacks fine details as well as local pose-dependent deformations and shading. Thus, no body renderings with loose clothes are demonstrated. While mesh-based techniques impose stronger human shape priors for higher quality results, they often assume water-tight clothing and a pre-captured template of the examined subject [10, 21, 20, 32]. This can lead to difficulties in capturing loosely-swinging garments or fine pose-dependent deformations. Furthermore, getting access to pre-captured templates can lead to additional assumptions about the scene (*e.g.,* that the target actor has to stay still for rigid structure-from-motion and brightness consistency for feature point matching) [21, 20].

3. Method

Our framework G_θ generates photorealistic videos of a target actor mimicking the motion of another source identity. Given only a single monocular RGB image sequence $I = (i_n)_{n=1}^N \in \mathbb{R}^{N \times 3 \times h \times w}$ containing N frames of width w and height h showing a target actor performing a rich set of motions, we extract an identity-independent 2D skeletal pose $P(i_n)$ for each frame and fit the network parameters θ to approximate the inverse mapping P^{-1} back to the original frames. Instead of performing the translation task within a single network, our framework can be thought of as a function composition of four generative neural networks (G_{ref}, G_{app}, G_{str}, G_{shp}) executed in a cascade fashion to progressively generate higher-level representations, as illustrated in Fig. 2. Next, we describe all individual framework components (Secs. 3.1-3.3), before providing details on the network architectures and implementation (Sec. 3.4).

3.1. Inputs

As inputs, our method takes the target image sequence I and a static image of the scene background $b \in \mathbb{R}^{3 \times h \times w}$ that is fused with the synthesized actor later on. From the given image sequence, we extract actor-independent pose representations $P : \mathbb{R}^{3 \times h \times w} \rightarrow H \times D$, where $H = \{0, 1\}^{C_1 \times h \times w}$ symbolizes the set of rasterized binary pose skeletons with C_1 input channels, and $D = \mathbb{R}^{C_2 \times h \times w}$ is the set of temporal derivatives for a single pose consisting of C_2 channels. Similar to recent performance cloning methods [6, 32], we apply an off-the-shelf pose estimator [4, 34, 40] that predicts 2D keypoints $k_n \in \mathbb{R}^{127 \times 2}$ for the body (including hands and face), to generate the skeleton $h_n \in H$ by connecting adjacent keypoints via binary lines. We further distribute $C_1 = 9$ limbs (face, head, torso, arms, legs, hands) over multiple channels, which helps the networks to distinguish between overlapping or symmetrical body parts.

However, estimating soft body deformations from a single pose remains highly ambiguous as the states depend on the temporal order of poses. Thus, we calculate a temporal pose context $d_n \in D$ as the first and second temporal derivatives of k_n with respect to the image index n in x and y direction, respectively ($C_2 = 2 \cdot 2 \cdot 9$), similar to velocities and accelerations in classical dynamics simulation:

$$k_n' = \left(\frac{\partial k_n}{\partial n}, \frac{\partial^2 k_n}{\partial n^2} \right). \tag{1}$$

We create rasterizations d_n from k_n' in the same way as for the pose skeletons, and linearly interpolate their values along the bones, similar to the depth representation of Shysheya *et al.* [32]. During the reenactment, we apply the same procedure to the source image sequence, but additionally perform pose normalization as described by Chan *et al.* [6] for inter-target appearance transfer before drawing the skeletons. Finally, our pose conditioning $P(i_n) = (h_n, d_n)$ is provided to our framework as the concatenation of the framewise pose skeleton and temporal context. We find that the described procedure results in a compact, yet expressive pose representation, while slightly outperforming a simple sliding window approach in our experiments.

3.2. Garment Conditioning

In our framework, we use two dedicated networks that pre-estimate the actor's shape and the internal gradient structure of clothing in 2D space to improve the consistent modeling of deforming garments and the fall of the folds.

3.2.1 Shape Estimator

As a first step, our shape estimator G_{shp} transforms the provided pose representations P into the actor's current silhouette

$$\tilde{s}_n^* = G_{shp}(P(i_n), \tilde{s}_{n-1}^*), \tag{2}$$

where $\tilde{s}_i^* \in \mathbb{R}^{j \times h \times w}$ is represented as semantic body-part segmentation with j labels. We extract a final segmentation mask $\tilde{s}_n \in ([\mathbb{N}]_1^j)^{h \times w}$ from our network by extracting the indices of maximum elements per pixel over the channel dimension.

As knowledge about the preceding shape is necessary to achieve temporal coherence of deforming garments, we design our networks in a recurrent fashion where each execution is conditioned on the previous output. Thus, G_{shp} is tasked to estimate a delta in the observed shape based on the current pose and the contextual temporal derivatives. The resulting segmentation mask is a significant prior for the final appearance network to produce visually-plausible renderings, and it is further used to segment the actor and parts of their clothing in later stages. For training, we use the human parsing method of Li *et al.* [18] trained on the
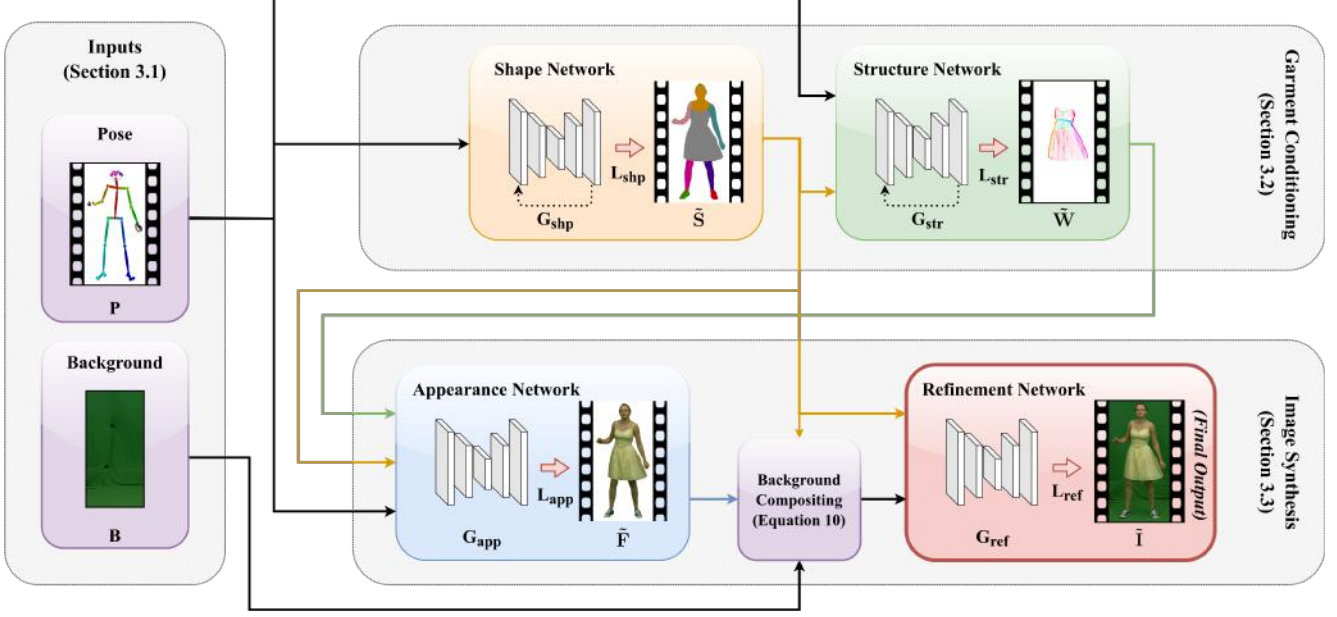
Figure 2. Framework Overview: Our framework synthesizes views of humans from a static background image and a pose sequence extracted from the target (training) or source (testing) image sequence. Therefore, we apply two consecutive networks to explicitly pre-estimate human body parsing (*yellow*) and the internal gradient structure of clothing (*green*). The resulting outputs are used to condition final image generation, which is separated into the foreground (*blue*) and global (*red*) components to handle loose clothing in a temporally coherent way. Dashed lines indicate recurrent networks that feedback the last output to generate the next prediction.

ATR dataset [19] to extract pseudo-ground-truth segmentation maps $S = (s_n)_{n=1}^n$ from I. This dataset provides $j = 18$ labels for all body limbs and common clothing, enabling our method to handle various clothing styles including shirts, pants, skirts, dresses and scarfs. We formulate our shape training loss L_{shp} as the cross-entropy between the network prediction and the training label:

$$L_{shp} = \mathbb{E}_{n \sim N} \log \left(\sum_j \exp(\tilde{s}_n^*(j)) \right) - \tilde{s}_n^*(s), \quad (3)$$

where $\tilde{s}_n^*(j)$ is the output channel for a given label j.

3.2.2 Structure Estimator

While our shape representation is sufficient prior to infer the appearance of solid body parts like arms and faces, it does not provide information about wrinkles and folds within the clothing, that are needed to generate temporally consistent shading and texture patterns. Thus, we apply a second recurrent network G_{str} that estimates the internal gradient structure $\tilde{w}_n \in \mathbb{R}^{2 \times h \times w}$ for clothing regions as indicated by the segmentation map $\tilde{s}_n$:

$$\tilde{w}_n = G_{str}(P(i_n), \tilde{s}_n, \tilde{w}_{n-1}). \quad (4)$$

We model the clothing structure as the pixelwise gradient direction and strength extracted from the responses of 32

oriented Gabor filters, similar to recent work on neural hair synthesis [36]. However, as a floating garment comprises sparser gradients than human hair, we do not only use the maximum filter responses to smooth the angle field, but append a normalized version to the final gradient directions to model the probability of a wrinkle or texture change at a specific location, as visualized in Fig. 1 (color and saturation encode the gradient direction and the strength, respectively). We train the structure estimator using the L_1 distance between estimated $\tilde{w}_n$ and ground-truth w_n structure using a mask for garment labels extracted from $\tilde{s}_n$:

$$L_{str} = \mathbb{E}_{n \sim N} \chi_C(\tilde{s}_n) |\tilde{w}_n - w_n|, \quad (5)$$

where C denotes the set of segmentation labels corresponding to the actor's clothing, and χ_C is the indicator function:

$$\chi_C(x) = \begin{cases} 1, & \text{if } x \in C, \\ 0, & \text{else.} \end{cases} \quad (6)$$

3.3. Image Synthesis

The second stage of our framework synthesizes the final output image based on the provided pose and garment conditionings. Again, we use two dedicated networks to independently generated the actor in the foreground and fuse it with the provided background image.

3.3.1 Appearance Network

Our first rendering network $\boldsymbol{G_{app}}$ takes provided pose as well as the pre-estimated shape $\tilde{S}$ and internal clothing structure $\tilde{W}$ to synthesize the actor's appearance $\tilde{f}_n \in \mathbb{R}^{3 \times h \times w}$:

$$\tilde{f}_n = \boldsymbol{G_{app}}(\boldsymbol{P}(i_n), \tilde{s}_n, \tilde{w}_n). \tag{7}$$

We train the appearance module using a combination of L_1 distance and perceptual reconstruction loss [14]. Again, an indicator function χ_B is used to mask the foreground pixels that are not assigned to the background label β:

$$\chi_B(x) = \begin{cases} 0, & \text{if } x = \beta, \\ 1, & \text{else.} \end{cases} \tag{8}$$

Thus, our appearance loss $\boldsymbol{L_{app}}$ reads

$$\boldsymbol{L_{app}} = \mathbb{E}_{n \sim N} \chi_B(\tilde{s}_n) \left(\lambda_r |\tilde{f}_n - i_n| + \lambda_p |\phi(\tilde{f}_n) - \phi(i_n)| \right), \tag{9}$$

where $\phi(\cdot)$ denotes feature maps extracted from different layers of a pre-trained VGG19 network [35], and λ_r, λ_p are free hyperparameters for weighting.

3.3.2 Refinement Network

Finally, we apply a shallow refinement network to fuse the foreground prediction with the provided scene background. For this, we first paste the generated foreground $\tilde{f}_n$ onto the static background image b using the masking function (8):

$$\tilde{i}_n^* = (\chi_B(\tilde{s}_n) \cdot \tilde{f}_n) + ((1 - \chi_B(\tilde{s}_n)) \cdot b). \tag{10}$$

Then, given foreground-background composition $\tilde{i}_n^*$ and actor segmentation $\tilde{s}_n$, our refinement network $\boldsymbol{G_{ref}}$ performs simple transition smoothing and shadow generation to produce the final output image $\tilde{i}_n$:

$$\tilde{i}_n = \boldsymbol{G_{ref}}(\tilde{i}_n^*, \tilde{s}_n). \tag{11}$$

We use a combination of structural and perceptual losses similar to our $\boldsymbol{L_{app}}$ (9), but over the entire image plane:

$$\boldsymbol{L_{ref}} = \mathbb{E}_{n \sim N} \lambda_r |\tilde{i}_n - i_n| + \lambda_p |\phi(\tilde{i}_n) - \phi(i_n)|. \tag{12}$$

In contrast to recent pose-to-video translation methods [1, 6], we do not apply an adversarial loss for our generator network as we do not want to hallucinate high frequency details in the clothing based on statistics from the data sequence I, but rather encourage the network to stick to the predicted structure layout.

3.4. Implementation Details

We implement our framework in Python using PyTorch 1.5 [27]. We intend to release the full implementation, including data sequences to reproduce the presented results.

Our four translation networks employ the local Pix2PixHD [38] generator architecture; we halve the number of residual blocks in our refinement network $\boldsymbol{G_{ref}}$ due to the similarity of the domains. We optimize our final loss

$$\boldsymbol{L_{total}} = \lambda_1 \boldsymbol{L_{seg}} + \lambda_2 \boldsymbol{L_{str}} + \lambda_3 \boldsymbol{L_{app}} + \lambda_4 \boldsymbol{L_{ref}} \tag{13}$$

with hyperparameters ($\lambda_1 = 0.5, \lambda_{2,3,4} = 1.0, \lambda_r = 0.1, \lambda_p = 0.9$) in a sequential way for 30 epochs using Adam optimization [16] (momentum $\beta_1 = 0.999$ and $\beta_2 = 0.5$). On a single Quadro RTX 8000 GPU, training converges after approximately four days for $N \approx 23K$ frames, while processing a single testing-frame takes $\sim$280 milliseconds at a resolution of 512×512 pixels.

4. Experiments

We compare our framework to several state-of-the-art approaches with subjects wearing different types of loose clothing with rich wrinkling and textures (Sec. 4.2), conduct a user study (Sec. 4.3) and perform an ablation study for individual framework stages (Sec. 4.4). Since our results improve upon the temporal consistency and fine texture details, they are best viewed in our supplementary video.

4.1. Datasets, Methods and Metrics

Existing datasets for the evaluation of recent appearance transfer methods usually comprise actors in monochromatic body-tight clothing, as loose garments are a frequent cause of visual artifacts and temporal irregularities. For our experiments, we capture new sequences of three subjects, each in up to three clothing styles with a length of seven to ten minutes at a resolution of 512×512 pixels, and select the last 5% of the frames for testing. Our dataset features various clothing styles including jeans, loose t-shirts, hoodies, and two dresses with rich wrinkle and texture patterns.

We compare our framework to current video-based appearance transfer approaches, *i.e.*, a general-purpose pix2pixHD network [38] that synthesizes an image from a rendering of the extracted pose, the Everybody-Dance-Now (EDN) method by Chan *et al.* [6], which adds a temporal discriminator and specialized face GAN, and Recycle-GAN [2] that performs direct unsupervised retargeting between two videos. Additionally, we test our approach on the dataset of Liu *et al.* [20] comprising various actors in tight clothes. All results were generated using either the official implementations or data directly provided by the authors. Moreover, we have explored pose editing using NHRR [30] and Liquid Warping GAN [22] but the results showed poor visual quality on our sequences. Kindly note that these methods predict pixel correspondences to the SMPL mesh [23] using DensePose [11], and hence struggle with handling loose clothes. We provide exemplary results showing limitations of mesh-based approaches in our

Figure 3. Exemplary results of our motion transfer framework. The first column of each block illustrates the driving sequence, while the last one shows the final result. They anchor the outputs for the different intermediate networks (shape, structure, and appearance). For clothing gradient structure, we encode the direction and confidence as color and saturation, respectively, similar to optical flow visualization.

supplementary material. Furthermore, the code of Textured Neural Avatars [32] is, unfortunately, not publicly available, and we were not able to obtain results on our sequences directly from the authors. For quantitative analysis, we utilize the widely-used SSIM [39], LPIPS [41] and FID [12] metrics that assess the pixel-space structural similarity, perceptual distance based on neural network features, and the Fréchet distance between two data collections, respectively.

4.2. Results and Comparisons

We show exemplary results of every network component in Fig. 3. As can be seen, our framework is capable of estimating pronounced body part segmentations, the corresponding wrinkle patterns and resulting target actor appearance for all types of complex clothing in our dataset, independent of the source actor's posture or appearance. We further compare our method to recent state-of-the-art approaches in Fig. 4. Therefore, we use a bidirectional reenactment scenario with two different types of clothing, *i.e.,* a loose monochromatic t-shirt with jeans, and a dress featuring a delicate arrangement of folds. While all of the tested methods are able to produce strong wrinkle patterns, our framework manages to achieve the most plausible temporal and spatial results due to the explicit handling of deforming garments. Please note that the temporal consistency can only be evaluated by watching our supplemental video. Also, we provide results and comparisons on the video sequences of Liu *et al.* [20] in our supplemental material.

We further assess the visual quality by numerically comparing our method to related pose-to-video translation approaches in Table 1. To calculate the metrics, we employ

Figure 4. Comparison to state-of-the-art motion transfer methods. We compare our method against pix2pixHD [38], EDN [6] and Recycle-GAN [2] on two sequences of an actress in a loose t-shirt and a dress. Please note that results of Recycle-GAN are only available at half of the original resolution, as training on full resolution produced highly implausible results.

the target actor sequences shown in the first column of Fig. 4 using the corresponding self-reenactment scenarios, where ground-truth images are available from the testing sets. Recycle-GAN [2] focuses on learning from unpaired data and is upper-bounded by paired translation techniques when trained on paired data. Hence, we exclude it from our quantitative analysis. Our method outperforms the current state of the art according to the SSIM and LPIPS metrics, which perform a direct structural and perceptual com-

Metric		pix2pixHD [38]	EdN [6]	ours
SSIM [39]	↑	0.9288	0.9043	**0.9358**
LPIPS [41]	↓	0.0401	0.0303	**0.0289**
FID [12]	↓	21.7724	**13.5421**	17.9573

Table 1. Quantitative comparison to related methods. We assess the quantitative quality according to SSIM, LPIPS and FID metrics in a self-reenactment scenario.

parison on a per-frame basis, indicating that the physical behaviour of clothing synthesized by our method better reflects ground truth. On the other hand, FID states that the pixel distribution generated by EDN is closer to the training set, presumably due to strong adversarial losses.

4.3. Perceptual Experiment

We find that quantitative metrics do not reveal all aspects of the visual video quality, and complement our evaluation by two user studies in which videos are graded according to human perception. We compare our method against recent image-based techniques using the video sequences shown in Fig. 4, including wide clothes with significant deformations. We also conduct an additional experiment on the dataset of Liu *et al.* [20] to evaluate on videos of actors in tight clothes. In both studies, we show the participants video reenactments of ∼20 seconds in length with results of two methods at a time and ask them which video looks the most realistic. To measure both the performance of each method and the agreement between participants, we follow the linked-paired comparison design [7]. We rank each method according to the number of times they are preferred over the rest and perform a significance test of the votes differences [31, 5]. The total number of votes a method received in each experiment is displayed in Table 2. As can be seen, our method significantly outperforms recent image-based pose-to-video-translation methods in terms of video realism for the clothing and is even preferred over Recycle-GAN, which uses a richer input domain for motion transfer. Our second experiment on the dataset of Liu *et al.* [20] further confirms that our monocular video-based approach also

Exp.	Method	Input	#Votes	Ranking
11	**Ours**	Video (Pose)	257	**1**
	Recycle-GAN [2]	Video (RGB)	194	**2**
	EDN [6]	Video (Pose)	143	**3**
	pix2pixHD [38]	Video (Pose)	54	**4**
2	**Liu** [20]	Textured mesh + Video (Pose)	103	**1**
	Ours	Video (Pose)	76	**2**
	EDN [6]	Video (Pose)	13	**3**

Table 2. Perceptual ranking of the compared methods for our user studies. Experiment 1 was conducted online with 54 participants with various backgrounds, while experiment 2 was in-situ with 16 CG/CV experts. The rankings are statistically significant.
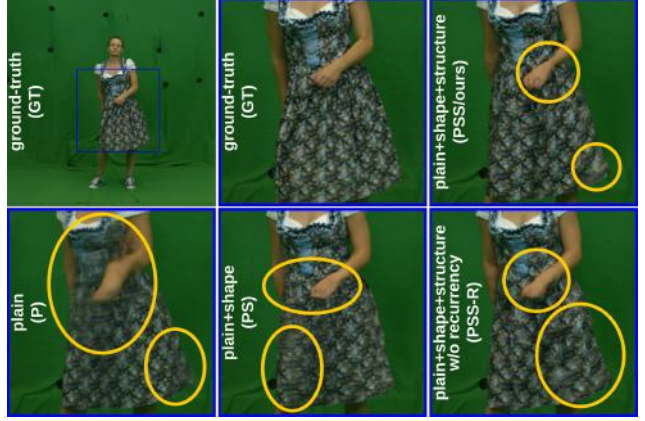

Figure 5. Ablation study evaluating the influence of our framework stages in a self-reenactment scenario. For an exemplary ground-truth frame, we show a plain appearance network conditioned on pose only (P), pose and shape (PS), and our full framework without (PSS-R) and with (PSS/ours) temporal recurrence.

achieves considerably more realistic results than the related EDN method when exchanging motion between actors in tight clothing. However, in that scenario, image-based approaches cannot quite reach the quality of priors provided by expensively recorded template meshes, which excel for tight and static shapes (Liu *et al.* [20]). We provide more details on the experimental setup and evaluation in our supplementary material.

4.4. Ablation Study

We next conduct an ablation study to assess the influence of different framework components. Therefore, we retrain our method on one of our most challenging sequences featuring a swinging dress with complicated texture patterns, while incrementally dropping the garment conditioning networks. In our study, we consider four framework variations, *i.e.,* the plain appearance network without additional inputs (P), the appearance conditioned on our shape estimates (PS) and our full framework including the internal structure (PSS/ours); additionally, we examine the effect of removing the recurrent back-feeding from all network modules (PSS-R). As shown in Fig. 5, we find that every component contributes to improving the final image quality.

While our plain pix2pix generator fails to produce sharp outlines of body parts and garments (P), providing a pre-estimated shape mask results in distinct transitions and more realistically-floating garments (PS). Adding information about the internal structure of the clothing further enhances the generation of high-frequency details within the given shape (PSS). This effect is also observable in quantitative evaluation using structural and perceptive metrics, as shown in Table 3. Furthermore, we observe that dropping the temporal back-feeding of previous network outputs

Figure 6. Applications: Our approach allows novel video editing through style transfer and wrinkle intensity manipulation. For style transfer, we use the source shape and/or structure while keeping the target appearance.

(PSS-R)—while achieving similar quality on a per-frame basis— significantly decreases temporal consistency, as can be seen in our supplemental video.

5. Application: Material Editing

On top of high-fidelity motion transfer, our framework offers an additional level of control over the image generation, as it incorporates explicit specifications of the actor's shape and garment structure. This enables editing and manipulating the generated videos. For actors wearing similar clothing, pseudo-ground-truth segmentation or structure maps extracted from the source sequence can significantly improve the final image quality and temporal consistency. Alternatively, after training multiple subjects wearing clothing of the similar type, garment conditioning networks can be interchanged to generate clothing with new combinations of shape, structure and appearance. Furthermore, manually-edited or completely handcrafted shape and structure patterns can be used to influence the physical behavior of clothing. Fig. 6 shows two examples of automatic and manual material and appearance editing. We change the style of the actress's dress and the actor's top while keeping the original appearance by applying the shape (left), or both shape and structure, from the source sequence and feed it to the appearance network (middle). Also, magnifying the second channel of structure estimate (gradient confidence map) can be used to magnify or minify the fall of the folds during

reenactment (right). While material property exchange is still currently limited to garments of comparable type or texture, the proposed approach is the first neural motion transfer method enabling controllable clothing composition for videos, which can pave the way for many applications in future (such as virtual try-on).

6. Discussion and Limitations

We have demonstrated high-fidelity results for multiple challenging human appearances with various deformation patterns of garments, and were able to manipulate several properties of the clothing during the reenactment. The ablative study has shown that all stages of our framework are justified and contribute to the final result. Although our method is a step forward in the neural rendering of humans, it has limitations and can provoke further research. First, the space of depictable poses is currently determined by the composition and length of the training sequences. Next, inaccuracies and missing 3D information from the pose keypoint estimator can result in incorrect occlusion maps and temporal behavior of garments. At the same time, the propagation of errors in the estimated human body parsing can lead to missing or disconnected body parts and unnatural transitions between the frames. These aspects can be improved in future by incorporating stronger constraints on the shape and body structure and enlarging the pose space by training on unpaired data.

7. Conclusion

We introduced a novel multi-stage framework for high-fidelity human motion transfer from monocular video. The tests showed that our method can perform qualitatively appealing motion transfer for different clothes, improving over previous image-based techniques in visual quality. The performance is also confirmed by a comprehensive perceptual study, indicating superior temporal and visual quality compared to current image-based state-of-the-art motion

Metric		P	PS	PSS-R	PSS (ours)
SSIM [39]	↑	0.8959	0.9045	0.9045	**0.9046**
LPIPS [41]	↓	0.044	0.036	0.036	**0.035**
FID [12]	↓	15.046	12.282	12.925	**12.176**

Table 3. Quantitative ablation analysis according to SSIM, LPIPS and FID metrics. We compare different versions of our framework in a self-reenactment scenario. The numbers indicate that every component contributes a structural and perceptual improvement to the synthesized output.

transfer approaches. All in all, we conclude that explicitly handling deformations of clothing in a temporally-coherent way is crucial for high-fidelity neural human motion transfer.

Acknowledgements

References

[1] Kfir Aberman, Mingyi Shi, Jing Liao, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. Deep video-based performance cloning. *Comput. Graph. Forum*, 38:219–233, 2018. 2, 5

[2] 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. 1, 2, 5, 6, 7, 11, 13

[3] David H. Brainard. The psychophysics toolbox. *Spatial Vision*, 10:433–436, 1997. 12

[4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(1):172–186, 2021. 3

[5] Susana Castillo, Tilke Judd, and Diego Gutierrez. Using eye-tracking to assess different image retargeting methods. In *Symposium on Applied Perception in Graphics and Visualization (APGV)*, page 7–14, 2011. 7

[6] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 3, 5, 6, 7, 11, 12, 13

[7] Herbert Aron David. *The method of paired comparisons*, volume 12 of *Griffin's statistical monographs & courses*. Charles Griffin, London, 1 1963. 7, 13

[8] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[9] Artur Grigorev, Artem Sevastopolsky, Alexander Vakhitov, and Victor Lempitsky. Coordinate-based texture inpainting for pose-guided human image generation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 12127–12136, 2019. 2

[10] Shanyan Guan, Shuo Wen, Dexin Yang, Bingbing Ni, Wendong Zhang, Jun Tang, and Xiaokang Yang. Human action transfer based on 3D model reconstruction. In *AAAI Conference on Artificial Intelligence*, pages 8352–8359, 2019. 2, 3

[11] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7297–7306, 2018. 5

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Neural Information Processing Systems (NIPS)*, pages 6626–6637, 2017. 6, 7, 8

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134, 2017. 2

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 694–711, 2016. 5

[15] Maurice G. Kendall and Bernard Babington Smith. On the method of paired comparisons. *Biometrika*, 31(3-4):324–345, 03 1940. 11

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[17] Mario Kleiner, David Brainard, and Denis Pelli. What's new in psychtoolbox-3? *Perception 36, European Conference on Visual Perception (ECVP) Abstract Supplement*, 2007. 12

[18] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *arXiv preprint arXiv:1910.09777*, 2019. 3, 11

[19] 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. 4, 11

[20] 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. 1, 2, 3, 5, 6, 7, 12, 13

[21] Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Hyeongwoo Kim, Florian Bernard, Marc Habermann, Wenping Wang, and Christian Theobalt. Neural rendering and reenactment of human actor videos. *ACM Transactions on Graphics (TOG)*, 2019. 2, 3

[22] Wen Liu, Zhixin Piao, Min Jie, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 5

[23] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 5

[24] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Neural Information Processing Systems (NIPS)*, 2017. 2

[25] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc van Gool, Bernt Schiele, and Mario Fritz. Disentangled person

image generation. *Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[26] Sylvain Paris, Will Chang, Oleg I. Kozhushnyan, Wojciech Jarosz, Wojciech Matusik, Matthias Zwicker, and Frédo Durand. Hair photobooth: Geometric and photometric acquisition of real hairstyles. *ACM Trans. Graph.*, 27(3), 2008. 2

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Neural Information Processing Systems (NIPS)*, pages 8024–8035, 2019. 5

[28] Egon S. Pearson and Herman O. Hartley. *Biometrika Tables for Statisticians*, volume 1. Cambridge University Press, 3 edition, 1966. 13

[29] Denis G. Pelli. The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, 10:437–442, 1997. 12

[30] Kripasindhu Sarkar, Dushyant Mehta, Weipeng Xu, Vladislav Golyanik, and Christian Theobalt. Neural re-rendering of humans from a single image. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 5

[31] Iwan Setyawan and Reginald L. Lagendijk. Human perception of geometric distortions in images. In *Security, Steganography, and Watermarking of Multimedia Contents VI*, volume 5306, pages 256 – 267, 2004. 7

[32] Aliaksandra Shysheya, Egor Zakharov, Kara-Ali Aliev, Renat Bashirov, Egor Burkov, Karim Iskakov, Aleksei Ivakhnenko, Yury Malkov, Igor Pasechnik, Dmitry Ulyanov, Alexander Vakhitov, and Victor Lempitsky. Textured neural avatars. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3, 6

[33] Aliaksandr Siarohin, Enver Sangineto, Stephane Lathuiliere, and Nicu Sebe. Deformable GANs for pose-based human image generation. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[34] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[36] 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. 2, 4, 11

[37] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Neural Information Processing Systems (NIPS)*, pages 1152–1164, 2018. 2

[38] 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. 1, 2, 5, 6, 7, 11, 13

[39] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 6, 7, 8

[40] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[41] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 6, 7, 8

[42] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *International Conference on Computer Vision (ICCV)*, 2017. 2

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

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

A. Garment Conditioning Representation and Visualization

Our current implementation uses the self-correction for human parsing method by Li *et al.* [18], that was trained on the ATR dataset [19]. 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 [18] provided by the authors[2]. 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.* [36] 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:

$$o_n = \arg\max_\Theta |(K_\Theta \otimes i_n)|,$$
$$c_n = \max_\Theta |(K_\Theta \otimes i_n)|,$$

$$(14)$$

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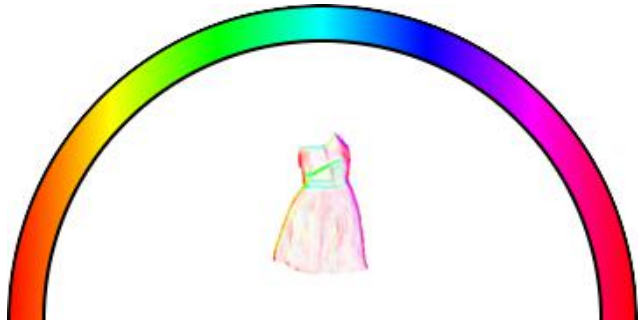
Figure I. Example of our garment structure visualization for one of the dresses shown in Fig. 1 from the main paper. We model the orientation as hue in HSV-color space, while each pixel's saturation corresponds to the local gradient confidence.

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. I.

B. 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 being quite subtle. More importantly, the quality of the results that we aim to measure cannot be represented on a linear scale [15], 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.

B.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 [6], pix2pixHD [38], and Recycle-GAN [2]) on some sequences of our data set (see the top part of Fig. II).
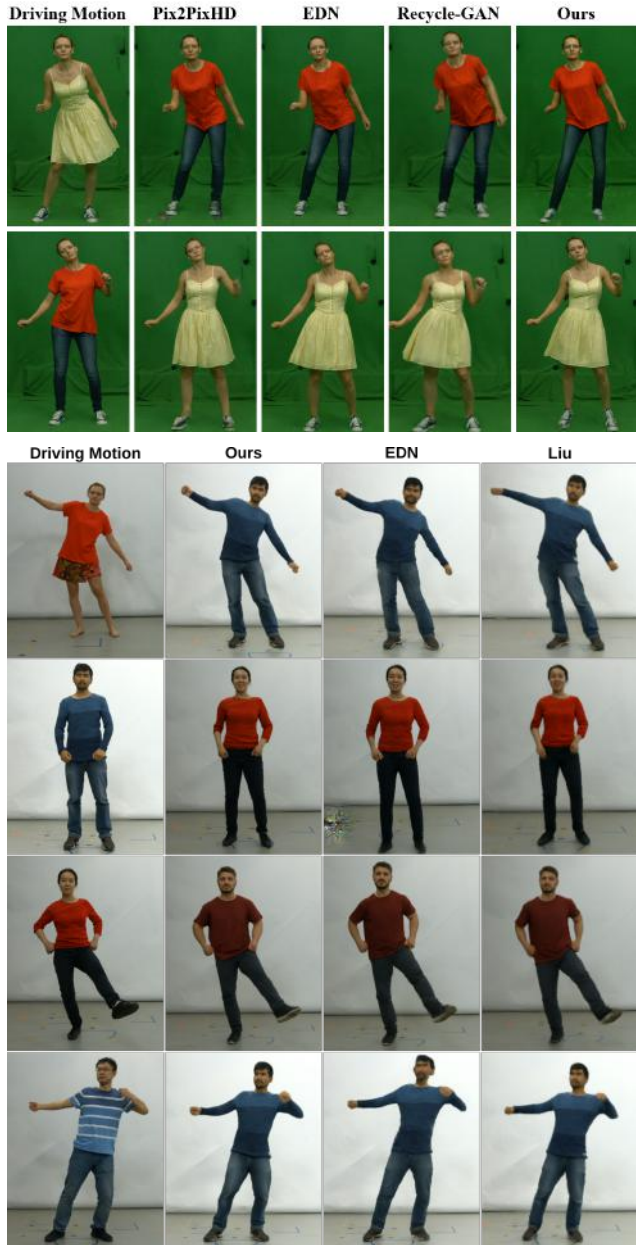
Figure II. Exemplary frames from the video sequences used as stimuli for our perceptual experiments. Top: *Exp. 1* conducted on our dataset. Bottom: *Exp. 2* on the dataset of Liu *et al.* [20].

The Second Experiment – Liu *et al.*'s Dataset. We also evaluated our technique on the dataset of Liu *et al.* [20] (see the second bottom part of Fig. II). Here, we trained two of the compared methods (EDN [6] 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 [20], EDN [6] 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.

B.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.

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 experimenter 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) [3, 29, 17]. 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.

B.3. Analysis

To measure not only the performance of each method but the agreement between participants, we followed the linked-paired comparison design [7]. 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 D for the first experiment ("*Exp. 1*") and Table E for the second experiment ("*Exp. 2*").

Exp. 1	Method	Input	#Votes	Ranking
	Ours	Video (Pose)	257	**1**
Q1	Recycle-GAN [2]	Video (RGB)	194	**2**
	EDN [6]	Video (Pose)	143	**3**
	pix2pixHD [38]	Video (Pose)	54	**4**
	Ours	Video (Pose)	243	**1**
Q2	Recycle-GAN [2]	Video (RGB)	205	**2**
	EDN [6]	Video (Pose)	135	**3**
	pix2pixHD [38]	Video (Pose)	65	**4**

Table D. *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 [20]	Textured mesh + Video (Pose)	103	**1**
	Ours	Video (Pose)	76	**2**
	EDN [6]	Video (Pose)	13	**3**

Table E. *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.

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 [7] we can derive R' from

$$P\left(W_{t,\alpha} \geq \frac{2R' - 0.5}{\sqrt{mt}}\right), \qquad (15)$$

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 [28]. In our case, $W_{4,0.01} = 4.405$ for *Exp. 1* (Table D) and $W_{3,0.01} = 4.125$ for *Exp. 2* (Table E). 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)"*).

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 F 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 F. 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.

C. Training Details

Our framework G_θ consist of four trainable components, namely G_{ref}, G_{app}, G_{str}, G_{shp}. For every network, we adapt the local Pix2PixHD [38] generator architecture, resulting in memory consumption and training times of $\sim 3.5\times$ the reported values for the original implementation (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 ~ 5GB VRAM and ~ 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, we feedback pseudo-ground-truth annotations during the first epoch, which encourages the network to expect reliable estimates of previous clothing deformations. We further synthesize the first frame of every video based on a black image (as no predecessor is available) and iteratively execute our networks on the initial input pose until the output converges towards a reasonable estimate (in practice, this takes around 20-30 iterations).