Learning Motion-Dependent Appearance for High-Fidelity Rendering of Dynamic Humans from a Single Camera

Jae Shin Yoon†,‡, Duygu Ceylan♯, Tuanfeng Y. Wang♯, Jingwan Lu♯, Jimei Yang♯, Zhixin Shu♯, Hyun Soo Park†

†University of Minnesota, ‡Adobe Research

Abstract

Appearance of dressed humans undergoes a complex geometric transformation induced not only by the static pose but also by its dynamics, i.e., there exists a number of cloth geometric configurations given a pose depending on the way it has moved. Such appearance modeling conditioned on motion has been largely neglected in existing human rendering methods, resulting in rendering of physically implausible motion. A key challenge of learning the dynamics of the appearance lies in the requirement of a prohibitively large amount of observations. In this paper, we present a compact motion representation by enforcing equivariance—a representation is expected to be transformed in the way that the pose is transformed. We model an equivariant encoder that can generate the generalizable representation from the spatial and temporal derivatives of the 3D body surface. This learned representation is decoded by a compositional multi-task decoder that renders high fidelity time-varying appearance. Our experiments show that our method can generate a temporally coherent video of dynamic humans for unseen body poses and novel views given a single view video.

1. Introduction

We express ourselves by moving our body that drives a sequence of natural secondary motion, e.g., dynamic movement of dress induced by dancing as shown in Figure 1. This secondary motion is the resultant of complex physical interactions with the body, which is, in general, time-varying. This presents a major challenge for plausible rendering of dynamic dressed humans in applications such as video based retargetting or social presence. Many existing approaches such as pose-guided person image generation [7] focus on static poses as a conditional variable. Despite its promising rendering quality, it fails to generate a physically plausible secondary motion, e.g., generating the same appearance for fast and slow motions.

One can learn the dynamics of the secondary motion from videos. This, however, requires a tremendous amount of data, i.e., videos depicting all possible poses and associated motions. In practice, only a short video clip is available, e.g., the maximum length of videos in social media (e.g., TikTok) are limited to 15-60 seconds. The learned representation is, therefore, prone to overfitting.

In this paper, we address the fundamental question of “can we learn a representation for dynamics given a limited amount of observations?”. We argue that a meaningful representation can be learned by enforcing an equivariant property—a representation is expected to be transformed in the way that the pose is transformed. With the equivariance, we model the dynamics of the secondary motion as a function of spatial and time derivative of the 3D body. We construct this representation by re-arranging 3D features in the canonical coordinate system of the body surface, i.e., the UV map, which is invariant to the choice of the 3D coordinate system. The UV map also captures the semantic meaning of body parts since each body part is represented by a UV patch. The resulting representation is compact and
discriminative compared to the 2D pose representations that often suffer from geometric ambiguity due to 2D projection.

We observe that two dominant factors significantly impact the physicality of the generated appearance. First, the silhouette of dressed humans is transformed according to the body movement and the physical properties (e.g., material) of the individual garment types (e.g., top and bottom garments might undergo different deformations). Second, the local geometry of the body and clothes is highly correlated, e.g., surface normals of T-shirt and body surface, which causes appearance and disappearance of folds and wrinkles. To incorporate these factors, we propose a compositional decoder that breaks down the final appearance rendering into modular subtasks. This decoder predicts the time-varying semantic maps and surface normals as intermediate representations. While the semantic maps capture the time-varying silhouette deformations, the surface normals are effective in synthesizing high quality textures, which further enables re-lighting. We combine these intermediate representations to produce the final appearance.

Our experiments show that our method can generate a temporally coherent video of an unseen secondary motion from novel views given a single view training video. We conduct thorough comparisons with various state-of-the-art baseline approaches. Thanks to the discriminative power, our representation demonstrates superior generalization ability, consistently outperforming previous methods when trained on shorter training videos. Furthermore, our method shows better performance in handling complex motion sequences including 3D rotations as well as rendering consistent views in applications such as free-viewpoint rendering. The intermediate representations predicted by our method such as surface normals also enable applications such as relighting which are otherwise not applicable.

2. Related Work

Two major rendering approaches have been used for high quality human synthesis: model based and retrieval based approaches. Model based approaches leverage the 3D shape models, e.g., 3DMM face model [4], that can synthesize a novel view based on the geometric transformation [11, 57]. Retrieval approaches synthesize an image by finding matches in both local and global shape and appearance [6, 31]. These are, now, combined with a deep representation to form neural rendering and generative networks.

Neural Rendering. As human appearances are modulated by their poses, it is possible to generate the high fidelity appearance by using a parametric 3D body model, e.g., deformable template models [35]. For example, some existing approaches [44, 67] learned a constant appearance of a person by mapping from per-vertex RGB colors defined on the SMPL body to synthesized images. Textured neural avatars [51] learned a person-specific texture map by projecting the image features to a body surface coordinate (invariant to poses) to model human appearances. These approaches, however, are limited to statics, i.e., the generated appearance is completely blind to pose and motion.

To model pose-dependent appearances, Liu et al. [34] implicitly learned the texture variation over poses, which allowed them to refine the initial appearance obtained from texture map through a template model. On the other hand, Raj et al. [47] explicitly learned pose-dependent neural textures. To further enhance the quality of rendering, person-specific template models [2] were used by incorporating additional meshes representing the garments [65]. However, none of these approaches are capable of modeling the time-varying secondary motion. Habermann et al. [19] utilized a motion cue to model the motion-dependent appearances while requiring a pre-learned person-specific 3D template model. Zhang et al. [68] proposed a neural rendering approach to synthesize the dynamic appearance of loose garments assuming a coarse 3D garment proxy is provided. In contrast, our method uses the 3D body prior to model the dynamic appearance of both tight and loose garments.

The requirement of the parametric model can be relaxed by leveraging flexible neural rendering fields. For instance, neural volumetric representations [39] have been used to model general dynamic scenes [15, 32, 56, 58, 64] and humans [41, 42] using deformation fields. Nonetheless, the range of generated motion is still limited. Recent methods learned the appearance of a person in the canonical space of the coarse 3D body template and employ skinning and volume rendering approaches to synthesize images [8, 43, 46]. Liu et al. [33] extended such approaches by introducing pose-dependent texture maps to model pose-dependent appearance. Modeling time-varying appearance induced by the secondary motion with such volumetric approaches is still the uncharted area of study.

Generative Networks. Generative adversarial learning enforces a generator to synthesize photorealistic images that are almost indistinguishable from the real images. For example, image-to-image translation can synthesize pose-conditioned appearance of a person by using various pose representations such as 2D keypoints [13, 36, 45, 49, 52, 55, 70], semantic labels [3, 12, 14, 20, 38, 54, 66], or dense surface coordinate parametrizations [1, 17, 40, 50]. Despite remarkable fidelity, these were built upon the 2D synthesis of static images at every frame, in general, failing to generate physically plausible secondary motion. To address this challenge, several works have utilized temporal cues either in training time [7] to enable temporally smooth results or as input signals [59, 60] to model motion-dependent appearances. Kappel et al. [25] modeled the pose-dependent appearances of loose garments conditioned on 2D keypoints by learning temporal coherence using a recurrent network. However, due to the nature of 2D pose representations, the
physicality of the generated motion is limited, e.g., our experiments show that the method works well mostly for planar motions but is limited in expressing 3D rotations. Wang et al. [62] is the closest to our work, which maps a sequence of dense surface parameterization to motion features that are used to synthesize dynamic appearances using StyleGAN [26]. In contrast, our method is built on a new 3D motion representation which shows superior discriminative power, consistently outperforming [62] in terms of generalizing to unseen poses as demonstrated in our experiments.

3. Method

Given a monocular video of a person in motion and the corresponding 3D body fit estimates, we learn a motion representation to describe the time-varying appearance of the secondary motion induced by body movement (Section 3.1). We propose a multitask compositional renderer (Section 3.2) that uses this representation to render the subject-specific final appearance of moving dressed humans. Our renderer first predicts two intermediate representations including time-varying semantic maps that capture garment specific silhouette deformations and surface normals that capture the local geometric changes such as folds and wrinkles. These intermediate representations are combined to synthesize the final appearance. We obtain the 3D body fits estimates from the input video using a new model-based tracking optimization (Supplementary material). The overview of our rendering framework is shown in Figure 2.

3.1. Equivariant 3D Motion Descriptor

We cast the problem of human rendering as learning a representation via a feature encoder-decoder framework:

$$f = E(p), \quad A = D(f),$$

where an encoder $E$ takes as an input a representation of a posed body, $p$ (e.g., 2D sparse or dense keypoints or 3D body surface vertices), and outputs per-pixel features $f$ that can be used by the decoder $D$ to reconstruct the appearance $A \in [0, 1]^{w \times h \times 3}$ of the corresponding pose where $w$ and $h$ are the width and height of the output image (appearance). We first discuss how $E$ can be modeled to render static appearance, then introduce our 3D motion descriptor to render time-varying appearance with secondary motion effects.

Learning a representation from Equation (1) from a limited amount of data is challenging because both encoder and decoder need to memorize every appearance in relation to the corresponding pose, $A \leftrightarrow p$. To address the data challenge, one can use an equivariant geometric transformation, $W$, such that a feature is expected to be transformed in the way that the body pose is transformed:

$$E(Wx) = WE(x).$$

where $x$ is an arbitrary pose. A naive encoder that satisfies this equivariance learns a constant feature $f_0$ by warping any $p$ to a neutral pose $p_0$:

$$f_0 = E(W^{-1}p_0) = \text{const.}, \quad A = D(Wf_0),$$

where $p = Wp_0$. Figure 3(a) and (b) illustrate cases where $W$ is defined as image warping in 2D or skinnyng in 3D respectively. $f$ can be derived by warping $p$ to the T-pose, $W^{-1}p_0$ of which feature can be warped back to the posed feature before decoding, $D(WWE(W^{-1}p))$. Since $f_0$ is constant, the encoder $E$ does not need to be learned. One can only learn the decoder $D$ to render a static appearance.

To model the time-varying appearance for the secondary motion that depends on both body pose and motion, one can extend Equation (3) to encode the spatial and temporal gradients as a residual feature encoding:

$$f = E(p, \frac{\partial p}{\partial x}, \frac{\partial p}{\partial t}) \approx E(p) + E_\Delta \left(\frac{\partial p}{\partial x}, \frac{\partial p}{\partial t}\right),$$

$$\iff f_0 = E(W^{-1}p_0) + E_\Delta \left(W^{-1}\frac{\partial p}{\partial x}, W^{-1}\frac{\partial p}{\partial t}\right).$$

Figure 2. The overview of our human rendering pipeline. Given a set of time-varying 3D body meshes $\{P_1, ..., P_{t-1}\}$ obtained from a monocular input video, we aim to synthesize high-fidelity appearance of a dressed human. We learn an effective 3D pose and motion representation by recording the surface normal $N^t$ of the posed 3D mesh at time $t$ and the body surface velocity $V^t$ over several past times in the spatially aligned UV space. We define an encoder $E_\Delta$ which is designed to reconstruct 3D motion descriptors $f_\Delta$ that encode the spatial and temporal relation of the 3D body meshes. Given a target 3D body configuration, we project $f_\Delta$ onto the image space which are then utilized by our compositional networks ($D_s$ and $D_a$) to predict a shape with semantic labels, surface normal, and final appearance.
where $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial t}$ are the spatial and temporal derivatives of the posed body, respectively. The spatial derivatives essentially represent the pose corrective deformations \cite{30, 35}. The temporal derivatives denote the body surface velocity which results in secondary motion. Since these spatial and temporal gradients are no longer constant, one needs to learn an encoder $E_\Delta$ to encode the residual features.

In this paper, we use a 3D representation of the posed body lifted from an image by leveraging recent success in single view pose reconstruction \cite{9, 24, 28}. Hence, spatial and temporal derivatives of the body pose correspond to the surface normals and body surface velocities, respectively:

$$f_{3D} = E_\Delta(W^{-1}N, W^{-1}V), \quad A = D(\Pi W f_{3D}),$$  \hspace{1cm} (5)

where $N = \frac{\partial p}{\partial x} \in \mathbb{R}^{m \times 3}$ is the 3D surface normal, and $V = \frac{\partial p}{\partial t} \in \mathbb{R}^{m \times 3}$ represents the instantaneous velocities of the $m$ vertices in the body surface. We model the geometric transformation function $W$ to warp an arbitrary 3D pose $p$ to a canonical representation, $p_0$. We record $f_{3D}$ in a spatially aligned 2D positional map, specifically the UV map of the 3D body mesh where each pixel contains the 3D information of a unique point on the body mesh surface. This enables us to leverage 2D CNNs to apply local convolution operations to capture the relationship between neighboring body parts \cite{37}. Therefore, $f_{3D} \in \mathbb{R}^{m \times d}$ called 3D motion descriptor is the feature defined in the UV coordinates where $d$ is the dimension of the per-vertex 3D feature. $f = \Pi W f_{3D}$ is the projected 3D feature in the image coordinates where $\Pi$ is a coordinate transformation that transports the features defined in the UV space to the image plane via the dense UV coordinates of the body mesh.

The key advantage of the 3D motion descriptor over commonly used 2D sparse \cite{25} or dense \cite{62} keypoint representations is discriminativity. Consider a toy example of a person rotating his body left to right multiple times. Given one cycle (i.e., $0$-$T$) of such a motion as input (Figure 4(a)), assume we want to synthesize the appearance of the person performing the repetitions of the same motion (i.e., cycles $T$-$5T$). As a proof of concept, we model $D$ using a nearest neighbor classifier to retrieve the relevant image patches (top two patches) for each body part from the reference motion based on the correlation of the motion descriptors as shown in Figure 4(c). Due to the inherent depth ambiguity, multiple 3D motion trajectories yield the same 2D projected trajectory \cite{23}. Hence, the 2D motion descriptors using sparse (Figure 4(b), top) and dense (Figure 4(b), middle) 2D keypoints confuse the directions of out-of-plane body rotation, resulting in erroneous nearest neighbor retrievals as shown in Figure 4(d). Furthermore, 2D representations entangle the viewpoint and pose into a common feature. This not only avoids a compact representation (e.g., the same body motion maps to different 2D trajectories with respect to different viewpoints and yields different features) but also suffers from occlusions in the image space. In our example in Figure 4(c), top and bottom, the upper arm is occluded in portions of the input video denoted with the purple blocks, hence no reliable local motion descriptors can be computed at those time instances. In contrast, our 3D motion descriptor is highly discriminative, which does not confuse the direction of body rotations, resulting in accurate image patch retrievals.

### 3.2. Multitask Compositional Decoder

Given 3D motion features, the decoder $D$ still needs to learn to generate diverse and plausible secondary motion, which is prone to overfitting given a limited amount of data. We integrate the following properties that can mitigate this challenge. (1) Composition: we design the decoder using a composition of modular functions where each modular function is learned to generate physically and semantically meaningful intermediate representations. Learning each modular function is more accurate than learning...
an end-to-end decoder as a whole as shown in our ablation study (Section 4.1); (2) Multi-task: each intermediate representation receives its own supervision signals resulting in multi-task learning. The motion features, \( f_{3D} \), are shared by all intermediate modules resulting in a compact representation; (3) Recurrence: each module is modeled as an autoregressive network, which allows learning the dynamical patterns rather than memorizing the pose-specific appearance.

Our decoder is a composition of two modular functions:

\[
D = D_a \circ D_s,
\]

where \( D_s \) and \( D_a \) are the functions that generate the shape with semantic maps and the appearance.

\( D_s \) learns the dynamics of the 2D shape:

\[
\tilde{s}_t = D_s(\tilde{s}_{t-1}; \tilde{f}_t)
\]

where \( \tilde{f}_t = \Pi W_t f_{3D} \) is the projected features onto the image at time \( t \), and \( \tilde{s}_t \in \{0, \cdots, L\}^{w \times h} \) is the predicted shape with semantics where \( L \) is the number of semantic categories. In our experiments we set \( L = 7 \) (background, top clothing, bottom clothing, face, hair, skin, shoes).

\( D_a \) learns the dynamics of appearance given the shape and 3D motion descriptor:

\[
\hat{A}_t, \hat{n}_t = D_a(\hat{A}_{t-1}, \hat{n}_{t-1}; \tilde{s}_t, \tilde{f}_t),
\]

where \( \hat{A}_t \in \mathbb{R}^{w \times h \times 3} \) and \( \hat{n}_t \in \mathbb{R}^{w \times h \times 3} \) are the generated appearance and surface normals at time \( t \).

We learn the 3D motion descriptor as well as the modular decoder functions by minimizing the following loss:

\[
\mathcal{L} = \sum_{P, A \in D} \mathcal{L}_a + \lambda_s \mathcal{L}_s + \lambda_n \mathcal{L}_n + \lambda_p \mathcal{L}_p + \lambda_g \mathcal{L}_g,
\]

where \( \mathcal{L}_a, \mathcal{L}_s, \mathcal{L}_n, \mathcal{L}_p, \mathcal{L}_g \) are the appearance, shape, surface normal, perceptual similarity, and generative adversarial losses, and \( \lambda_s, \lambda_n, \lambda_p, \lambda_g \) are their weights, respectively. We set \( \lambda_s = 10, \lambda_n = \lambda_p = 1 \) and \( \lambda_g = 0.01 \) in our experiments. \( D \) is the training dataset composed of the ground truth 3D pose \( P \) and its appearance \( A \).

\[
\begin{align*}
\mathcal{L}_a(P, A) &= \| \hat{A} - A \|, \\
\mathcal{L}_s(P, A) &= \| \hat{s} - S(A) \|, \\
\mathcal{L}_n(P, A) &= \| \hat{n} - N(A) \|, \\
\mathcal{L}_p(P, A) &= \sum_i \| VGG_i(\hat{A}) - VGG_i(A) \|, \\
\mathcal{L}_g(P, A) &= E_{S(A), \hat{A}}[\log(D^*(S(A), \hat{A})] + E_{S(A), A}[\log(1 - D^*(S(A), \hat{A})],
\end{align*}
\]

where \( \hat{A}, \hat{s}, \hat{n} \) are the generated appearance, shape, and surface normal, respectively. \( S \) and \( N \) are the shape [16] and surface normal estimates [22], and \( VGG \) is the feature extractor that computes perceptual features from conv-i-2 layers in VGG-16 networks [53]. \( D^* \) is the PatchGAN discriminator [21] that validates the plausibility of the synthesized image conditioned on the shape mask.
3.3. Model-based Monocular 3D Pose Tracking

While there has been significant improvements in monocular 3D body estimation [9, 28, 29], we observe that predicting accurate and temporally coherent 3D body sequences is still challenging, which inhibit to reconstruct high-quality 3D motion descriptors. Hence, we devise a new optimization framework that learns a tracking function to address this challenge. We describe the full methodology and evaluation of our 3D tracking pipeline in the supplementary materials.

4. Experiments

We validate the performance of our method across various examples and perform extensive qualitative and quantitative comparisons with previous work.

Implementation Details. We utilize the Adam optimizer [27] to train our model with a learning rate of $1 \times 10^{-3}$. Given an input video ($\sim 10^5$ frames), we train our model for roughly 72 hours using 4 NVIDIA V100 GPUs using a batch size of 4. Our motion features are learned from the body surface normals in the current frame and the body surface velocities in the past $t = 10$ frames. The features are recorded in a UV map of size $128 \times 128$. We synthesize final renderings and surface normal maps of size $512 \times 512$. We implement our model in Pytorch and utilize the Pytorch3D differentiable rendering layers [48]. Details of the network designs and 3D tracking pipeline are given in the supplementary materials.

The coordinate transformation that transports the motion features from the UV space to the image plane, i.e., II in Equation (5), can be implemented either by using image-based dense UV estimates [18] or by directly rendering the UV coordinates of the 3D body fits. To provide a fair comparison with previous work which also utilize dense UV estimates, we use the former option. When demonstrating our method on applications where we do not have corresponding ground truth frames to estimate dense UV maps (e.g., novel viewpoint synthesis), we use the latter option.

Baselines. We compare ours to four prior methods that focused on synthesizing dressed humans in motion. 1) Everybody dance now (EDN) [7] uses image-to-image translation to synthesize human appearance conditioned on 2D keypoints and uses a temporal discriminator to enforce plausible dynamic appearance. 2) Video-to-video translation (V2V) [61] is a sequential video generator that synthesizes high-quality human renderings from 2D keypoints and dense UV maps where the motion is modeled with optical flow in the image space. 3) High-fidelity motion transfer (HFMT) [25] is a compositional recurrent network which predicts plausible motion-dependent shape and appearance from 2D keypoints. 4) Dance in the wild (DIW) [62] synthesizes dynamic appearance of humans based on a motion descriptor composed of time-consecutive 2D keypoints and dense UV maps. We evaluate only on foreground by removing the background synthesized by the methods EDN, V2V, and DIW using a human segmentation method [16]. HFMT predicts a foreground mask similar to ours. In the supplementary material, we also compare our method to the 3D based approach [8].
for neural avatar modeling from a single camera, which explicitly reconstruct the geometry of animate human.

Datasets. We perform experiments on video sequences that demonstrate a wide range of motion sequences and clothing types, which include non-trivial secondary motion. Specifically, we select three dance videos (e.g., hip-hop and salsa) from YouTube and one sequence from prior work [25] that shows a female subject in a large dress. We also capture two custom sequences showing a male and a female subject respectively performing assorted motions (e.g., walking, running, punching, jumping etc.) including 3D rotations.

Metrics. We measure the quality of the synthesized frames with two metrics: 1) Structure similarity (SSIM) [63] compares the local patterns of pixel intensity in the normalized luminance and contrast space. 2) Perceptual distance (LPIPS) [69] evaluates the cognitive similarity of a synthesized image to ground truth by comparing their perceptual features extracted from a deep neural network. We evaluate the temporal plausibility by comparing the perceptual change across frames [10]: tLPIPS = \|LPIPS(s_t, s_{t-1}) − LPIPS(g_t, g_{t-1})\| where s and g are the synthesized and ground truth images.

4.1. Evaluation

Comparisons We provide quantitative evaluation in Table 1 and show qualitative results in Figure 5 (see Supplementary Video). Similar to our method, we train each baseline for roughly 72 hours until convergence. Both qualitative and quantitative results show that sparse 2D keypoint based pose representation used in EDN is not as effective as other baselines or our method. HFMT is successful in modeling dynamic appearance changes for mostly planar motions (i.e., MPI sequence), but shows inferior performance in remaining sequences that involve 3D rotations. This is due to the depth ambiguity inherent in sparse 2D keypoint based representation. While V2V performs well in terms of quantitative numbers, it suffers from significant texture drifting issues as shown in Figure 5, second row. We speculate that this is due to the errors in the optical flow estimation, especially in case of loose clothing, which is used as a supervisory signal. DIW uses dense UV coordinates to model the dynamic appearance changes of loose garments and is the strongest baseline. While it performs consistently well, we observe that the performance gap between DIW and our method increases for motion segments consisting of 3D rotations. This gap is magnified when the testing motion deviates from the training data. In Figure 6, we plot how the perceptual error changes along the motion similarity between the training and testing data which is computed by NCC between two sets of the time-varying 3D meshes similar to Figure 4. We observe bigger increase in the error for DIW as testing frames deviate more from the training data.

We next perform further comparisons with DIW evaluating the generalization ability of each approach. Generalization An effective motion representation that encodes the dynamic appearance change of dressed humans should be discriminative to distinguish all possible deformations induced by a pose transformation given the current state of the body and the garments. In order to compare the discriminative power of our motion descriptor and the dense keypoint based representation proposed by DIW, we evaluate how well each representation generalizes to unseen poses. Specifically, we train each model using only 10% of the original training sequences by subsampling the training frames while ensuring training and testing pose sequences are sufficiently distinct. Considering the reduced amount of data, we limit the training time to 24 hours for both approaches. As shown in Table 2, the performance gap be-
We presented a method to render the dynamic appearance of a dressed human given a reference monocular video. Our method utilizes a novel 3D motion descriptor that encodes the time varying appearance of garments to model effects such as secondary motion. Our experiments show that our 3D motion descriptor is effective in modeling complex motion sequences involving 3D rotations. Our descriptor also demonstrates superior discriminator power compared to state-of-the-art alternatives enabling our method to better generalize to novel poses.

While showing impressive results, our method still has limitations. Highly articulated hand regions can appear blurry, hence refining the appearance of such regions with specialized modules is a promising direction. Our current model is subject specific, extending different parts of the model, e.g., 3D motion descriptor learning, to be universal is also an interesting future direction.

Broader Impact While our goal is to enable content creation for tasks such as video based motion retargeting or social presence, there can be some misuse of our technology to fabricate fake videos or news. We hope that parallel advances in deep face detection and image forensics can help to mitigate such concerns.

Acknowledgement

We would like to thank Julien Philip for providing useful feedback on our paper draft. Jae Shin Yoon is supported by Doctoral Dissertation Fellowship from University of Minnesota. This work is partially supported by NSF CNS-1919965.
References


Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. NeurIPS, 2019. 2
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 CVPR, 2018. 6, 7
Tuanfeng Y. Wang, Duygu Ceylan, Krishna Kumar Singh, and Niloy J. Mitra. Dance in the wild: Monocular human animation with neural dynamic appearance synthesis, 2021. 3, 4, 6, 7, 8
Donglai Xiang, Fabian Andres Prada, Timur Bagautdinov, Weipeng Xu, Yuan Dong, He Wen, Jessica Hodgins, and Christoph Lassner. Space-time neural irradiance fields for free-viewpoint video. In CVPR, 2021. 2
Jae Shin Yoon, Lingjie Liu, Vladislav Golyanik, Kripasindhu Sarkar, Hyun Soo Park, and Christian Theobalt. Pose-guided rendering of humans from a single image. 2
Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. NeurIPS, 2019. 2
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 CVPR, 2018. 6, 7
Tuanfeng Y. Wang, Duygu Ceylan, Krishna Kumar Singh, and Niloy J. Mitra. Dance in the wild: Monocular human animation with neural dynamic appearance synthesis, 2021. 3, 4, 6, 7, 8
Jae Shin Yoon, Lingjie Liu, Vladislav Golyanik, Kripasindhu Sarkar, Hyun Soo Park, and Christian Theobalt. Pose-guided rendering of humans from a single image. 2
human animation from a single image in the wild. In CVPR, 2021. 2

