Flow Guided Transformable Bottleneck Networks for Motion Retargeting

Jian Ren Menglei Chai Oliver J. Woodford* Kyle Olszewski Sergey Tulyakov
Snap Inc.
{jren, mchai, kolszewski, stulyakov}@snap.com

Abstract

Human motion retargeting aims to transfer the motion of one person in a "driving" video or set of images to another person. Existing efforts leverage a long training video from each target person to train a subject-specific motion transfer model. However, the scalability of such methods is limited, as each model can only generate videos for the given target subject, and such training videos are labor-intensive to acquire and process. Few-shot motion transfer techniques, which only require one or a few images from a target, have recently drawn considerable attention. Methods addressing this task generally use either 2D or explicit 3D representations to transfer motion, and in doing so, sacrifice either accurate geometric modeling or the flexibility of an end-to-end learned representation. Inspired by the Transformable Bottleneck Network, which renders novel views and manipulations of rigid *objects, we propose an approach based on an implicit volumetric representation of the image content, which can then be spatially manipulated using volumetric flow fields. We address the challenging question of how to aggregate information across different body poses, learning flow fields that allow for combining content from the appropriate regions of input images of highly* non-rigid *human subjects performing complex motions into a single implicit volumetric representation. This allows us to learn our 3D representation solely from videos of moving people. Armed with both 3D object understanding and end-to-end learned rendering, this categorically novel representation delivers state-of-the-art image generation quality, as shown by our quantitative and qualitative evaluations.*

1. Introduction

Retargeting human body motion — transferring motion from a "driving" image or video of one subject (the *source*) to another subject (the *target*), using one or more reference images of the target subject in an arbitrary pose — has received a great deal of attention in recent years, due to numerous practical and entertaining applications in content generation [62, 8]. Such applications include transferring sophisticated athletic techniques or dancing performances to untrained celebrities for special effects for cinema and television; creating amusing performances for one's friends or acquaintances for sheer entertainment; and creating plausible motion sequences from photos or videos depicting famous and important political figures (including historical figures who may no longer be alive to perform such actions) for the creation of plausible full-body "deepfake" videos. However, retaining the target subject's identity while rendering them in novel, unseen poses is highly challenging, and the state-of-the-art is still far from plausible.

Many approaches to this task learn to render a specific person [1, 7, 8, 13, 20, 26, 49, 64, 63, 66, 67, 73] conditioned on the desired pose. This requires a large number of training frames of that person, and incurs substantial training time that must be repeated per each new subject. By contrast, in the few-shot setting, addressed in this work, only a few reference images of the target are available, and video generation from those images should be fast (*i.e.*, requiring no subject-specific training). To overcome the lack of data for a given subject, many other techniques [4, 29, 32, 34, 40, 43, 44, 46] leverage existing human body models [2, 22, 36] to construct an approximate representation of the subject that can then be manipulated and rendered. While the 3D nature of these representations often leads to improved performance over their purely 2D counterparts [3, 38, 39, 51, 55], their explicit nature, which faces the limitations of capturing salient details with standard human body models, also leads to reduced modeling power and therefore fidelity. Large variations in the clothing (*e.g.*, dresses or jackets that do not conform to the body shape), body type, or hair of the source and target subjects, for example, cannot easily be represented with standard models that only represent the body itself.

In this work we attain more flexible and expressive modeling power by exploiting a representation that allows for 3D modeling and manipulation, and yet is fully implicit, *i.e.* it can be fully learned, even though we use no explicit

*Work done while at Snap Inc.

ground-truth 3D information such as meshes or voxel grids as supervision. Recently, just such a representation, the Transformable Bottleneck Network (TBN) [45], has been shown to produce excellent results on novel view synthesis of rigid objects. In that work, image content is encoded into an implicit volumetric representation (the "bottleneck"), in which each of the encoded features in this volume correspond to the local structure and appearance of the corresponding spatial region in the volume depicted in the image. However, while it requires no 3D supervision, it is trained using multi-view datasets of *rigid* objects depicted from multiple viewpoints to produce implicit volumes that can then be rigidly transformed to produce novel views of the depicted content corresponding to changes in viewpoint.

We build upon this approach to address the challenge of performing motion retargeting for *non-rigid* humans (for which multiple images of a given subject may be available, but in dramatically different poses). In doing so, we address several challenges: how to aggregate volumetric features from images with changes in camera and body pose, and how to learn this aggregation from videos without explicit 3D or camera pose supervision. With such an implicit representation, to synthesize a novel pose, we achieve *non-rigid* implicit volume aggregation and manipulation by learning a 3D flow to resample the 3D body model from input images captured with the subject performing various poses or under different viewpoints. To allow for expressing large-scale motion while retaining fine-grained details in the synthesized images, we propose a multi-resolution scheme in which image content is encoded, transformed and aggregated into bottlenecks of different resolution scales.

As we focus on transferring motion between human subjects, our network pipeline is designed and trained specifically to extract and manipulate the foreground of the encoded images, with a separate network for extracting and compositing the background with the synthesized result. Our training scheme employs techniques and loss functions precisely designed for the challenging task of producing plausible motion retargeting without 3D supervision or the use of explicit 3D models, *e.g*. making use of specialized training techniques to teach the network to synthesize plausible results when no ground-truth images corresponding to the applied spatial manipulation is available. We thus avoid the limitations of explicit body representations [32, 34], which may lead to unrealistic results due to the limited reconstruction accuracy and mesh precision. Furthermore, it allows for learning directly from real 2D images and videos without requiring the tedious and cumbersome collection of copious high-fidelity 3D data [29, 44, 14].

In our experiments, we demonstrate that our approach qualitatively and quantitatively outperforms state-of-the-art approaches to human motion transfer, despite the few images used for inference, and even allows for plausible motion transfer when using only a single image of the target.

In summary, our key contributions are:

- A novel set of neural network architectures for performing implicit volumetric human motion retargeting, which exploits the power of 3D human motion modeling while avoiding the limitations of standard 3D human body modeling techniques.
- A framework to train these networks to attain high-fidelity human motion transfer using only a few example images of the target subject performing various poses, *without* requiring target-specific training.
- Evaluations demonstrating our few-shot approach outperforms state-of-the-art alternatives both quantitatively and qualitatively, even those requiring training models for each new subject with substantial training data.

2. Related Work

Video-to-Video Generation. Existing works on video-to-video generation can synthesize high-quality human motion videos using conditional image generation [7, 20, 64]. Chan *et al*. [8] apply pre-computed human poses from driving videos as input for novel view and pose generation of a target person. Along this line, several works improve the synthesis quality through additional input signals [13], pose augmentation and pre-processing [49, 66, 67, 73], and temporal coherence [1, 63]. However, a long recorded video and person-specific training are required for each target person, limiting the scalability of such methods. Our work targets a few-shot scenario, with just a handful of source images available, as discussed below.

Few-Shot Motion Retargeting. To address scalability, others train generic models that can use just a few images of the target subject in arbitrary poses at test time to synthesize novel images with given poses. While high-quality results have been achieved for face animation [15, 16, 48, 69, 70], animating bodies remains challenging. Some methods adapt the video-to-video approach to a few-shot setting, using an extra network to generate identity-specific weights for image generation network [62], or adapting a pre-trained network to new subjects [30]. Another class of methods train networks conditioned on an image of the source identity, and an explicit representation of target pose [38, 39, 12]. More recent works exploit multi-stage networks to improve quality [55, 19, 10], in particular using 2D spatial transformers [3, 50, 51] or deformation [58] to warp the source image into the target pose, or synthesizing images using attention [61, 74]. However, such purely 2D methods struggle to capture the complex motions generated by 3D shapes and transformations. Several works use explicit 3D representations, exploiting off-the-shelf human pose [2] and shape [22] inference networks and body meshes [36]. DensePose [2] is used to unwrap ap-

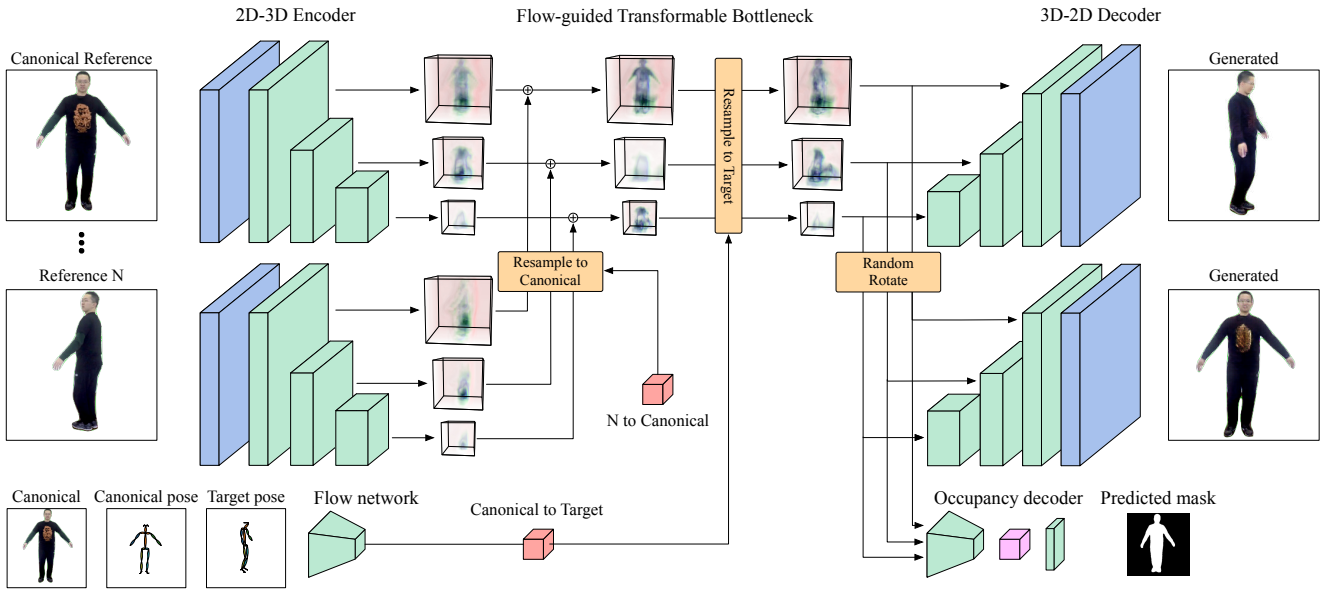

Figure 1. Overall pipeline. Given N reference images (left), we extract implicit volumetric representations at multiple scales using our 2D-3D encoder. Our flow network (bottom left) computes the 3D warping field used to warp these representations from the canonical pose (the first source image, to which the other source image feature volumes are aligned) to the target pose. Our 3D-2D decoder then synthesizes the subject in the target pose (top right), while transformations such as rotations can be applied to synthesize novel views of the subject in this pose (middle right). Our occupancy decoder (bottom right) improves the bottleneck's spatial disentanglement and thus allows for better motion retargeting. It computes an occupancy volume used to indicate which regions of the volume are occupied by the subject, which is then decoded into a 2D foreground segmentation mask.

pearance from the source image(s) into a canonical texture map, which is then inpainted and re-rendered in the target pose [29, 44, 14]. Other works use 3D meshes to compute 2D flow fields to warp features from source to target images [32, 34], with shallow encoders and decoders at each end to reduce warping artifacts. Finally, several works use body models and standard rendering techniques for clothing transfer [4, 40, 43, 46]. However, such explicit representations generally produce synthetic-looking results. We use an *implicit* 3D representation on this task for the first time, unlocking several benefits: the motion representation is more flexible, no 3D supervision or prior is needed, the image decoder can refine the output easily, and multiple reference images can be used to improve synthesis quality.

Unsupervised Motion Retargeting. Unsupervised methods learn retargeting purely from videos [57, 28, 24, 60, 47], forgoing motion supervision and therefore also class-specific motion representations. Siarohin *et al*. learn unsupervised keypoints in order to warp object parts into novel poses [56, 57]. Lorenz *et al*. [37] show that body parts can also be represented by unsupervised learning, which also helps to disentangle body pose and shape [11]. Nevertheless, without the benefit of explicit pose information from a human detection model, these methods struggle to generate good results for challenging driving poses. We therefore use a high quality, off-the-shelf, pose detection model [5] to facilitate the generation process.

Implicit and Volumetric 3D Representations. The flexibility of implicit 3D content representations [54, 6, 27], which obviate the use of explicit surfaces, *e.g*. discrete triangle meshes, make them amenable to recent learning-based approaches that extract the information needed to render or reconstruct scenes directly from images. In [59], an object-specific network is trained from many calibrated images to extract deep features embedded in a 3D grid that are used to synthesize novel views of the object. [42] perform novel view synthesis on complex scenes by aggregating samples from a trained network which, given a point in space and a view direction, provides an opacity and color value indicating the radiance towards the viewer from this point. This approach was later extended to handle unconstrained, large-scale scenes [41], or to use multiple radiance field functions, stored in sparse voxel fields to better capture detailed scenes with large empty regions [33]. However, these approaches share the limitation that the networks are trained on many images with known camera poses for a specific scene, and thus cannot be used to perform 3D reconstruction or novel view synthesis from new images without re-training. Furthermore, many images from different viewpoints are required to train these networks to learn to sufficiently render points between these images. Other recent efforts use 3D feature grids [23] or pixel-wise implicit functions [52, 53, 31] to infer dynamic shapes from one or more images, but these require synthetic ground-truth 3D

geometry for supervision, which limits their applicability to unconstrained real-world conditions. In [45], an encoder-decoder framework is used to extract a volumetric implicit representation of image content that is spatially disentangled in a manner that allows for novel view synthesis, 3D reconstruction and spatial manipulation of the encoded image content. However, while it allows for performing non-rigid manipulations of the image content after training, it must be trained using a multi-view dataset of *rigid* objects, making it unsuitable for human motion retargeting. In our work we develop an enhanced implicit volumetric representation and multi-view fusion techniques to address these concerns.

3. Method

At the heart of our approach to few-shot human body motion retargeting is an implicit volumetric representation of the shape and appearance of the subject depicted in the input images. In the following sections, we first describe this representation, then outline the network architectures employing it to perform foreground motion transfer and composition into the background environment. Finally, we discuss the training techniques and loss functions employed to train these networks to perform flexible motion retargeting while still achieving high-fidelity image synthesis results. The overall architecture is illustrated in Figure 1.

3.1. Implicit Volumetric Representation

Given an input image, our encoder extracts a feature volume or "bottleneck" $f \in \mathbb{R}^{N \times D \times H \times W}$, where D, H, and W are the depth, height, and width of the encoded volumetric grid used to represent the image content. Each cell in the grid contains an N-dimensional feature vector describing the structure and appearance of the local image content corresponding to that region of the volume. The depth dimension of this volume is aligned with the view direction of the camera, while the height and width correspond to those of the input image. The feature volume may be passed directly to the image decoder to synthesize an image corresponding to the input (in which case it acts as an auto-encoder), or it may be spatially manipulated in a manner corresponding to the desired transformation of the image content. Such manipulations include rigid transformations corresponding to camera viewpoint changes, or non-rigid manipulations corresponding to the subject's body motion.

Given the encoded bottleneck f and a dense *flow field* $\mathrm{T}_{s \to t} \in \mathbb{R}^{3 \times D \times H \times W}$ encoding a 3D coordinate per cell in the transformed bottleneck f' pointing to the original volume f that corresponds to the mapping from pose s to t, we employ a sampling layer $\mathrm{S} : f, \mathrm{T}_{s \to t} \to f'$ to perform trilinear sampling to produce f'. This flexible sampling mechanism enables a wide variety of spatial manipulations, from rigid transformations for novel view synthesis (*e.g.* simulating camera pose change by rotating the bottleneck) to more complex non-rigid changes (*e.g.* raising the arms of the subject while keeping the rest of the body stationary).

While the flow field $\mathrm{T}_{s \to t}$ for rigid camera viewpoint changes can be easily computed given the relative camera transformation between s and t, non-rigid body pose transformations can be much more complicated. It requires the flow field to have the appropriate sampling location in f for each cell in f' that semantically corresponds to the desired 3D motion of the content extracted from the input image.

Our training process, as described below, in which both rigid viewpoint and non-rigid pose transformations are used, achieves the *spatial disentanglement* required to infer the appropriate flow field and employ it to guide the desired spatial transformation of the image content.

3.2. Flow-Guided Volumetric Resampling

Network Architecture Overview. Our human body synthesis branch G_{fg} has three major components: the encoder, which consists of a 2D encoding network $\mathrm{Enc}_{\mathrm{2D}}$ and a 3D encoding network $\mathrm{Enc}_{\mathrm{3D}}$; the decoder, which consists of a 3D decoding network $\mathrm{Dec}_{\mathrm{3D}}$ and a 2D decoding network $\mathrm{Dec}_{\mathrm{2D}}$; and the flow-guided transformable bottleneck network F. A source reference image $\mathbf{x}_{\mathrm{fg}}$, containing only the foreground region from the original image $\mathbf{x}$ (obtained using a pre-trained segmentation model [9]), is passed through the encoder to obtain the 3D feature representation f as $f = \mathrm{Enc}_{\mathrm{3D}}(\mathrm{Enc}_{\mathrm{2D}}(\mathbf{x}_{\mathrm{fg}}))$. To warp the features, we estimate the flow using the flow network F with inputs of $\mathbf{x}_{\mathrm{fg}}$ and the source/target poses $\mathbf{p}_s, \mathbf{p}_t$, which are obtained with the off-the-shelf pose detection network [5, 2]. For simplicity, we define this operation as $\mathrm{F}_{s \to t}(f) \coloneqq \mathrm{S}(f, \mathrm{F}(\mathbf{x}_{\mathrm{fg}}, \mathbf{p}_s, \mathbf{p}_t))$, and generate the synthesized foreground frame using the flow field from F as:

$$\hat{\mathbf{y}}_{\mathrm{fg}} = \mathrm{Dec}_{\mathrm{2D}}(\mathrm{Dec}_{\mathrm{3D}}(\mathrm{F}_{s \to t}(f))). \tag{1}$$

Multi-Resolution Bottleneck. To increase the fine-scale fidelity of the synthesized images while retaining the global structure of the target subject, we adopt a multi-resolution representation, with implicit volumetric representations at multiple resolutions, as seen in Figure 1. We employ skip connections between each 3D encoder and 3D decoder. Therefore, the encoder produces 3D features such that $f = \{f_1, f_2, \ldots, f_m\}$, where m is the number of resolutions. The skip connections use 3D warping, given the estimated flow, to map content to from the correct region in the encoded bottleneck to the one to be decoded, rather than the direct connections used in prior art [17]. Thus, for each feature f_i obtained from the 3D encoder, the 3D decoder receives as input $\mathrm{F}_{s \to t}(f_i)$.

Multi-View Aggregation. Given that a single input image only contains partial information for the depicted human body, we allow for the aggregation of information, represented with our 3D features, from multiple views to im-

prove the synthesized image quality. To accelerate the inference, we take the first source image as the canonical body pose (though this may actually be the subject in any natural pose), and aggregate features from all other images to this pose. In this way, we only perform the aggregation once during inference. More formally, for a total of N source images, the aggregated feature f_a is represented as:

$$f_a = \frac{1}{N}(f_1 + \sum_{i=2}^{N} \mathrm{F}_{i\to 1}(f_i)), \tag{2}$$

where f_i is the 3D representation from the image i and $\mathrm{F}_{i\to 1}$ is the warping flow from the image i to the first image.

3.3. Background Modeling

The background environment is modeled through a separate network G_{bg}, to which the source images are fed. Specifically, G_{bg} estimates a background image $\hat{\mathbf{y}}_{\mathrm{bg}}$ and a confidence map $\hat{\mathbf{w}}$, in which high confidence indicates the foreground, *i.e*. the depicted person, while low confidence indicates the background. The synthesized image $\hat{\mathbf{y}}$ is:

$$\hat{\mathbf{y}} = \hat{\mathbf{w}} \cdot \hat{\mathbf{y}}_{\mathrm{fg}} + (1 - \hat{\mathbf{w}}) \cdot \hat{\mathbf{y}}_{\mathrm{bg}}, \tag{3}$$

where $\cdot$ denotes component-wise multiplication of the confidence map with the color channels of the synthesized image, and $\hat{\mathbf{y}}_{\mathrm{bg}}$ indicates the synthesized background image.

3.4. Training

Retargeting Supervision. We use a conditional discriminator D_{fg} to determine whether the synthesized foreground image is real or fake. The concatenation of the input image, 2D source/target poses, and target image is sent to the discriminator and we apply the following adversarial loss:

$$\begin{aligned}\mathcal{L}_{\mathrm{D}_{\mathrm{fg}}} = \mathbb{E}_{\mathbf{p}_t,\mathbf{y}_{\mathrm{fg}}}[\log \mathrm{D}_{\mathrm{fg}}(\mathbf{p}_t, \mathbf{y}_{\mathrm{fg}})] + \\ \mathbb{E}_{\mathbf{x}_{\mathrm{fg}},\mathbf{p}_s,\mathbf{p}_t}[\log(1 - \mathrm{D}_{\mathrm{fg}}(\mathbf{p}_t, \mathrm{G}_{\mathrm{fg}}(\mathbf{x}_{\mathrm{fg}}, \mathbf{p}_s, \mathbf{p}_t)))],\end{aligned} \tag{4}$$

where $\mathbf{y}_{\mathrm{fg}}$ is the foreground region from the real image. The generator is trained to minimize this objective, while the discriminator is trained to maximize it.

Similarly, we have a discriminator D_{bg} that works on the full synthesized with foreground and background:

$$\begin{aligned}\mathcal{L}_{\mathrm{D}_{\mathrm{bg}}} = \mathbb{E}_{\mathbf{p}_t,\mathbf{y}}[\log \mathrm{D}_{\mathrm{bg}}(\mathbf{p}_t, \mathbf{y})] + \\ \mathbb{E}_{\mathbf{x},\mathbf{p}_s,\mathbf{p}_t}[\log(1 - \mathrm{D}_{\mathrm{bg}}(\mathbf{p}_t, \mathrm{G}_{\mathrm{bg}}(\mathbf{x}, \mathbf{p}_s, \mathbf{p}_t)))],\end{aligned} \tag{5}$$

where $\mathbf{y}$ denotes the real image.

We also use a perceptual loss [21] $\mathcal{L}_{\mathrm{vgg}}$ between the real and generated images to improve fidelity of the generated images for both the foreground and background:

$$\mathcal{L}_{\mathrm{per}} = \mathcal{L}_{\mathrm{vgg}}(\mathbf{y}_{\mathrm{fg}}, \hat{\mathbf{y}}_{\mathrm{fg}}) + \mathcal{L}_{\mathrm{vgg}}(\mathbf{y}, \hat{\mathbf{y}}). \tag{6}$$

Additionally, we measure the reconstruction quality of each of the N source images using the aggregated bottleneck. The reconstructed image is generated as $\hat{\mathbf{x}} = \mathrm{Dec}_{\mathrm{2D}}(\mathrm{Dec}_{\mathrm{3D}}((f)))$, with no flow required for this auto-encoding. The reconstruction loss $\mathcal{L}_{\mathrm{recon}}$ is:

$$\mathcal{L}_{\mathrm{recon}} = \frac{1}{N}(\sum_{i=1}^{N} \|\hat{\mathbf{x}} - \mathbf{x}\|_1). \tag{7}$$

Mask Supervision. We also leverage mask supervision, using the foreground masks, to better supervise the implicit 3D representation modeling. Similar to previous work [45], we introduce an occupancy decoder to obtain the estimated mask directly from the 3D features. Considering that we have multi-resolution bottlenecks in our architecture, we apply multiple occupancy decoders to get the mask from each of the bottlenecks. Specifically, given the occupancy decoder as $\mathrm{Dec}_{\mathrm{occ}}$, the estimated mask $\hat{\mathbf{w}}_i$ for the i-th 3D representation is given as $\hat{\mathbf{w}}_i = \mathrm{Dec}_{\mathrm{occ}}(f_i)$. The mask loss $\mathcal{L}_{\mathrm{mask}}$ is thus defined as follows:

$$\mathcal{L}_{\mathrm{mask}} = \frac{1}{N}(\sum_{j=1}^{N} \frac{1}{m}(\sum_{i=1}^{m} \left\| \hat{\mathbf{w}}_i^j - \mathbf{w}_i^j \right\|_1)), \tag{8}$$

where $\mathbf{w}^j$ is the mask obtained from j-th source reference image using the pre-trained segmentation network.

Unsupervised Random Rotation Supervision. We further introduce an unsupervised training technique to help the network learn implicit volumetric representations for the encoded subject with appropriate spatial structure. Specifically, we apply a random rotation around the vertical axis of the volume to the encoded bottleneck and enforce the corresponding synthesized image to be indistinguishable from the ground-truth views. The magnitude of this rotation is sampled from a uniform distribution $r \sim U(-180°, 180°)$. The synthesized image is thus $\hat{\mathbf{y}}_r = \mathrm{Dec}_{\mathrm{2D}}(\mathrm{Dec}_{\mathrm{3D}}(\mathrm{F}_{s\to r}(f)))$, which should contain a novel view of the subject performing the same pose as in the source image, where $\mathrm{F}_{s\to r}$ is the flow field defined by the rigid transformation between the source pose s and the random pose r. However, since there is no ground-truth image corresponding to each random rigid transformation, we introduce a discriminator $\mathrm{D}_{\mathrm{rot}}$ to match the distribution between real images and $\hat{\mathbf{y}}_r$ in an unsupervised manner. We provide the discriminator with the concatenation of the generated image and the original source to better maintain the source identity. We employ an adversarial loss to enforce the rotation constraint on the foreground region as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{D}_{\mathrm{rot}}} = \mathbb{E}_{\mathbf{x}_{\mathrm{fg}},\mathbf{y}_{\mathrm{fg}}}[\log \mathrm{D}_{\mathrm{rot}}(\mathbf{x}_{\mathrm{fg}}, \mathbf{y}_{\mathrm{fg}})] + \\ \mathbb{E}_{\mathbf{x}_{\mathrm{fg}},\mathbf{p}_s,r}[\log(1 - \mathrm{D}_{\mathrm{rot}}(\mathbf{x}_{\mathrm{fg}}, \mathrm{G}_{\mathrm{fg}}(\mathbf{x}_{\mathrm{fg}}, \mathbf{p}_s, r)))].\end{aligned} \tag{9}$$

Full Objective. The full training objective for the entire

Table 1. Quantitative results on the iPER [34] dataset. We compare our method with existing works using the SSIM and LPIPS.

Method	SSIM↑	LPIPS↓
PG2 [38]	0.854	0.135
SHUP [3]	0.832	0.099
DSC [58]	0.829	0.129
LiquidGAN [34]	0.840	0.087
Ours	**0.868**	**0.086**

motion retargeting network is thus given as follows:

$$\mathcal{L}_{\mathrm{T}} = \lambda_{\mathrm{fg}}\mathcal{L}_{\mathrm{D_{fg}}} + \lambda_{\mathrm{bg}}\mathcal{L}_{\mathrm{D_{bg}}} + \lambda_{\mathrm{per}}\mathcal{L}_{\mathrm{per}} + \lambda_{\mathrm{recon}}\mathcal{L}_{\mathrm{recon}} + \lambda_{\mathrm{mask}}\mathcal{L}_{\mathrm{mask}} + \lambda_{\mathrm{rot}}\mathcal{L}_{\mathrm{D_{rot}}}, \quad (10)$$

where λ_{fg}, λ_{bg}, λ_{per}, λ_{recon}, λ_{mask}, and λ_{rot} are hyper-parameters to control the weight of each loss function.

4. Experiments

In this section, we evaluate our method on two different datasets, and show qualitative and quantitative comparisons with recent state-of-the-art efforts on human motion generation. For additional results, including video sequences, please consult the supplemental material.

4.1. Experimental Setting

Datasets. We adopt two datasets for experiments:

- **iPER**. The first dataset, known as the Impersonator (iPER) dataset, was recently collected and released by LiquidGAN [34], and serves as a benchmark dataset for human motion animation tasks. There are 206 videos with a total of $241,564$ frames in the dataset. Each video depicts one person performing different actions. We follow the protocol in previous work, splitting the training and testing set in the ratio of 8:2. Compared with other datasets [35, 72, 68], iPER includes human subjects with widely varying shape, height and appearance, and performing diverse motions.
- **Youtube-Dancing.** We create this dataset by collecting dancing videos from Youtube. It includes 675 dancing videos with $4,063,453$ frames in total for training, and 200 videos with $12,965$ frames for testing. The videos are recorded in unconstrained, in-the-wild environments, and consist of more challenging and diverse motion patterns compared with the iPER dataset.

Implementation Details. We normalize all images to the range $[-1, 1]$ for training, and train the networks to synthesize images with a resolution of 256×256. We apply the Adam optimizer [25] with a mini-batch size of 32 for training. Following previous work [64], we apply multi-scale discriminators, where each discriminator accepts images with the original resolution, 256×256 and images with half this resolution, 128×128. The hyper-parameters in Eqn. 10 are set to 1, except for $\lambda_{\mathrm{per}} = 10$.

Table 2. Quantitative results on the Youtube-Dancing dataset. We report the SSIM, LPIPS, and FID for both methods.

Method	SSIM↑	LPIPS↓	FID↓
FewShot-V2V [62]	0.770	0.144	48.39
Ours	**0.787**	**0.131**	**18.82**

Table 3. Ablation analysis of our proposed architecture on the iPER dataset. Our Full method achieves results that are superior to all other variants.

Method	SSIM↑	LPIPS↓
w/o Multi-Resolution TBN	0.870	0.051
w/o Skip-Connections	0.856	0.059
w/o Random Rotations	0.876	0.047
Full	**0.878**	**0.045**

Table 4. Analysis of the use of different numbers of reference images from the target subject for image generation on the iPER dataset. The image quality improves with our method as more reference images are used.

Method	SSIM↑	LPIPS↓
One reference image	0.878	0.045
Two reference images	0.879	0.045
Three reference images	0.881	0.044
Four reference images	**0.882**	**0.043**

Evaluation Metrics. We use three widely-adopted metrics to evaluate the image synthesis quality. The Structural Similarity (SSIM) [65] index measures the structural similarity between synthesized and real images. The Learned Perceptual Similarity (LPIPS) [71] measures the perceptual similarity between these images. The Fréchet Inception Distance (FID) [18] calculates the distance between two distributions of real and synthesized images, and is commonly used to quantify the fidelity of the synthesized images.

4.2. Comparisons and Results

We first show the experimental results on the iPER dataset. Following previous work [34], we use three reference images with different degrees of occlusion from each video to synthesize other images. We report both the SSIM and LPIPS for our work and existing studies, including PG2 [38], SHUP [3], DSC [58], and LiquidGAN [34] in Table 1. As shown in the table, we achieve better results using both metrics than state-of-the-art works, indicating higher synthesized image quality with our method. Additionally, we provide qualitative results in Figure 2. We show an example source reference image from a target subject, and the generated sequence using the driving pose information from the ground-truth images. As depicted, our method can generate realistic images with diverse motion patterns.

We then perform these experiments on our collected

Figure 2. Qualitative results on the iPER dataset. The leftmost columns show the reference images. For each example, we show both ground-truth (GT) driving images (odd rows), and the images generated by our method using the pose from the GT images (even rows). Our method can generate realistic images from diverse motion patterns.

Youtube-Dancing dataset. We compare our method with the most recent work, FewShot-V2V [62], which can generate motion retargeting videos using one or a few images. For evaluation, we use the first frame from each testing video as a source reference frame and using poses from the other frames to generate images. Besides the SSIM and LPIPS, we also follow FewShot-V2V [62] and report the FID scores. The results are summarized in Table 2. As can be seen, our method achieves better results than FewShot-V2V [62] on each of the three metrics. We also provide sample qualitative results in Figure 3. We show the source reference image from a target subject, and the generated short sequence from our method and FewShot-V2V using the pose from the ground-truth images. Compared with FewShot-V2V, our method generates more realistic images with fewer artifacts, and the identity and the texture from the target subject are much better preserved.

4.3. Ablation Study

Architecture and Training Technique Analysis. We conduct an ablation analysis of our network architectures and training strategies. For this experiment, we analyze the effects of each component used to generate the human body, *i.e*. the foreground region. We thus remove the background using an image segmentation model [9], and only compute evaluation metrics on the generated foreground region. We use one source reference image from a target subject and preform experiments with the following setting on iPER:

- *w/o Multi-Resolution TBN.* Instead of using multi-resolution bottlenecks, we use a single resolution TBN to encode the implicit 3D representation.
- *w/o Skip-Connections.* The skip-connections (implemented via flow warping) between the encoded and decoded 3D bottlenecks are removed.
- *w/o Random Rotations.* We remove the unsupervised random rotation supervision applied to the encoded TBN and its associated loss function, defined in Eqn. 9.

The quantitative results, including the SSIM and LPIPS, are shown in Table 3. We can see that each component is beneficial and that our full method (*Full*) achieves the best results.

Multi-View Aggregation Analysis. As our method can leverage multiple reference images from the target to perform multi-view aggregation, we conduct experiments to

Figure 3. Qualitative results on the Youtube-Dancing dataset. The first and fifth column show one reference image from a target person. For each target person, we show short sequences, including the ground-truth driving images, and the images generated by our and FewShot-V2V [62] using the pose information from the driving images. As seen here, our approach clearly generates more realistic results.

analyze the effect of the number of reference images on the synthesis result using the iPER dataset. The experimental results, presented in Table 4, demonstrate that using more reference images improves the quality of the synthesized human bodies.

5. Conclusion

Our approach to few-shot human motion retargeting exploits advantages of 3D representations of human body while avoiding limitations of the more straightforward prior methods. Our implicit 3D representation, learned via spatial disentanglement during training, avoids pitfalls of standard geometric representations such as dense pose estimations or template meshes, which are limited in their expressive capacity and for which it is impossible to obtain accurate ground-truth in unconstrained conditions. However, it allows for 3D-aware motion inference and image content manipulation, and attains state-of-the-art results on challenging motion retargeting benchmarks. Though we require 2D human poses, our approach could be extended to allow for more general motion retargeting for images of articulated animals given their 2D poses.

References

[1] Kfir Aberman, Mingyi Shi, Jing Liao, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. Deep video-based performance cloning. In *Computer Graphics Forum*, volume 38, pages 219–233. Wiley Online Library, 2019.

[2] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018.

[3] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8340–8348, 2018.

[4] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5420–5430, 2019.

[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018.

[6] J. C. Carr, R. K. Beatson, J. B. Cherrie, T. J. Mitchell, W. R. Fright, B. C. McCallum, and T. R. Evans. Reconstruction and representation of 3d objects with radial basis functions. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '01, page 67–76, New York, NY, USA, 2001. Association for Computing Machinery.

[7] Menglei Chai, Jian Ren, and Sergey Tulyakov. Neural hair rendering. In *Eur. Conf. Comput. Vis.*, 2020.

[8] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A. Efros. Everybody dance now. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[9] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[10] Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. Soft-gated warping-gan for pose-guided person image synthesis. In *Advances in neural information processing systems*, pages 474–484, 2018.

[11] Patrick Esser, Johannes Haux, and Bjorn Ommer. Unsupervised robust disentangling of latent characteristics for image synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2699–2709, 2019.

[12] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8857–8866, 2018.

[13] Oran Gafni, Lior Wolf, and Yaniv Taigman. Vid2game: Controllable characters extracted from real-world videos. *arXiv preprint arXiv:1904.08379*, 2019.

[14] Artur Grigorev, Artem Sevastopolsky, Alexander Vakhitov, and Victor Lempitsky. Coordinate-based texture inpainting for pose-guided human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12135–12144, 2019.

[15] Kuangxiao Gu, Yuqian Zhou, and Thomas Huang. Flnet: Landmark driven fetching and learning network for faithful talking facial animation synthesis. *arXiv preprint arXiv:1911.09224*, 2019.

[16] Sungjoo Ha, Martin Kersner, Beomsu Kim, Seokjun Seo, and Dongyoung Kim. Marionette: Few-shot face reenactment preserving identity of unseen targets. *arXiv preprint arXiv:1911.08139*, 2019.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[18] 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 *NeurIPs*, 2017.

[19] Mohamed Ilyes Lakhal, Oswald Lanz, and Andrea Cavallaro. Pose guided human image synthesis by view disentanglement and enhanced weighting loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[20] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[21] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[22] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018.

[23] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. In *Adv. Neural Inform. Process. Syst.*, pages 365–376, 2017.

[24] Yunji Kim, Seonghyeon Nam, In Cho, and Seon Joo Kim. Unsupervised keypoint learning for guiding class-conditional video prediction. In *Advances in Neural Information Processing Systems*, pages 3814–3824, 2019.

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

[26] Markus Knoche, István Sárándi, and Bastian Leibe. Reposing humans by warping 3d features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1044–1045, 2020.

[27] Aaron Knoll. A survey of implicit surface rendering methods, and a proposal for a common sampling framework. In *Lecture Notes in Informatics (LNI), Proceedings - Series of the Gesellschaft fur Informatik (GI)*, pages 164–177, 01 2007.

[28] Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, Nicu Sebe, et al. Motion-supervised co-part segmentation. *arXiv preprint arXiv:2004.03234*, 2020.

[29] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-degree textures of people in clothing from a single image. In *2019 International Conference on 3D Vision (3DV)*, pages 643–653. IEEE, 2019.

[30] Jessica Lee, Deva Ramanan, and Rohit Girdhar. Metapix: Few-shot video retargeting. *arXiv preprint arXiv:1910.04742*, 2019.

[31] Ruilong Li, Yuliang Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li. Monocular real-time volumetric performance capture. *arXiv preprint arXiv:2007.13988*, 2020.

[32] Yining Li, Chen Huang, and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3693–3702, 2019.

[33] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *NeurIPS*, 2020.

[34] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Int. Conf. Comput. Vis.*, pages 5904–5913, 2019.

[35] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016.

[36] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.

[37] Dominik Lorenz, Leonard Bereska, Timo Milbich, and Bjorn Ommer. Unsupervised part-based disentangling of object shape and appearance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10955–10964, 2019.

[38] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Adv. Neural Inform. Process. Syst.*, pages 406–416, 2017.

[39] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 99–108, 2018.

[40] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J Black. Learning to dress 3d people in generative clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6469–6478, 2020.

[41] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections, 2020.

[42] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[43] Aymen Mir, Thiemo Alldieck, and Gerard Pons-Moll. Learning to transfer texture from clothing images to 3d humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7023–7034, 2020.

[44] Natalia Neverova, Riza Alp Guler, and Iasonas Kokkinos. Dense pose transfer. In *Proceedings of the European conference on computer vision (ECCV)*, pages 123–138, 2018.

[45] Kyle Olszewski, Sergey Tulyakov, Oliver Woodford, Hao Li, and Linjie Luo. Transformable bottleneck networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7648–7657, 2019.

[46] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7365–7375, 2020.

[47] Albert Pumarola, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. Unsupervised person image synthesis in arbitrary poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8620–8628, 2018.

[48] Shengju Qian, Kwan-Yee Lin, Wayne Wu, Yangxiaokang Liu, Quan Wang, Fumin Shen, Chen Qian, and Ran He. Make a face: Towards arbitrary high fidelity face manipulation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10033–10042, 2019.

[49] Jian Ren, Menglei Chai, Sergey Tulyakov, Chen Fang, Xiaohui Shen, and Jianchao Yang. Human motion transfer from poses in the wild. *arXiv preprint arXiv:2004.03142*, 2020.

[50] Yurui Ren, Ge Li, Shan Liu, and Thomas H Li. Deep spatial transformation for pose-guided person image generation and animation. *IEEE Trans. Image Process.*, 29:8622–8635, 2020.

[51] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H Li, and Ge Li. Deep image spatial transformation for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7690–7699, 2020.

[52] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, 2019.

[53] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2020.

[54] Chen Shen, James F. O'Brien, and Jonathan R. Shewchuk. Interpolating and approximating implicit surfaces from polygon soup. *ACM Trans. Graph.*, 23(3):896–904, Aug. 2004.

[55] Chenyang Si, Wei Wang, Liang Wang, and Tieniu Tan. Multistage adversarial losses for pose-based human image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 118–126, 2018.

[56] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2377–2386, 2019.

[57] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Adv. Neural Inform. Process. Syst.*, December 2019.

[58] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3408–3416, 2018.

[59] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proc. CVPR*, 2019.

[60] Sijie Song, Wei Zhang, Jiaying Liu, and Tao Mei. Unsupervised person image generation with semantic parsing transformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2357–2366, 2019.

[61] Hao Tang, Song Bai, Li Zhang, Philip HS Torr, and Nicu Sebe. Xinggan for person image generation. *arXiv preprint arXiv:2007.09278*, 2020.

[62] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. In *Adv. Neural Inform. Process. Syst.*, 2019.

[63] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018.

[64] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[65] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004.

[66] Dongxu Wei, Xiaowei Xu, Haibin Shen, and Kejie Huang. Gac-gan: A general method for appearance-controllable human video motion transfer. *IEEE Transactions on Multimedia*, 2020.

[67] Zhuoqian Yang, Wentao Zhu, Wayne Wu, Chen Qian, Qiang Zhou, Bolei Zhou, and Chen Change Loy. Transmomo: Invariance-driven unsupervised video motion retargeting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5306–5315, 2020.

[68] Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. Dwnet: Dense warp-based network for pose-guided human video generation. *arXiv preprint arXiv:1910.09139*, 2019.

[69] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *European Conference on Computer Vision*, pages 524–540. Springer, 2020.

[70] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9459–9468, 2019.

[71] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[72] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.

[73] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara Berg. Dance dance generation: Motion transfer for internet videos. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[74] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2347–2356, 2019.