Liquid Warping GAN with Attention: A Unified Framework for Human Image Synthesis

Wen Liu, Zhixin Piao, Zhi Tu, Wenhan Luo, Lin Ma, and Shenghua Gao

Abstract—We tackle human image synthesis, including human motion imitation, appearance transfer, and novel view synthesis, within a unified framework. It means that the model, once being trained, can be used to handle all these tasks. The existing task-specific methods mainly use 2D keypoints (pose) to estimate the human body structure. However, they only express the position information with no abilities to characterize the personalized shape of the person and model the limb rotations. In this paper, we propose to use a 3D body mesh recovery module to disentangle the pose and shape. It can not only model the joint location and rotation but also characterize the personalized body shape. To preserve the source information, such as texture, style, color, and face identity, we propose an Attentional Liquid Warping GAN with Attentional Liquid Warping Block (AttLWB) that propagates the source information in both image and feature spaces to the synthesized reference. Specifically, the source features are extracted by a denoising convolutional auto-encoder for characterizing the source identity well. Furthermore, our proposed method can support a more flexible warping from multiple sources. To further improve the generalization ability of the unseen source images, a one/few-shot adversarial learning is applied. In detail, it firstly trains a model in an extensive training set. Then, it finetunes the model by one/few-shot unseen image(s) in a self-supervised way to generate high-resolution (512×512 and 1024×1024) results. Also, we build a new dataset, namely Impersonator (iPER) dataset, for the evaluation of human motion imitation, appearance transfer, and novel view synthesis. Extensive experiments demonstrate the effectiveness of our methods in terms of preserving face identity, shape consistency, and clothes details. All codes and dataset are available on https://impersonator.org/work/impersonator-plus-plus.html.

Index Terms—Human Image Synthesis, Motion Imitation, Appearance Transfer, Novel View Synthesis, Generative Adversarial Network, and One/Few-Shot Learning

1 Introduction

Human image synthesis aims to make believable and photo-realistic images of humans, including motion imitation [1], [2], [3], appearance transfer [4], [5] and novel view synthesis [6], [7]. It has vast potential applications in character animation, re-enactment, virtual clothes try-on, movie or game making, etc. Given a source human image and a human reference image, i) the goal of motion imitation is to generate an image with the texture from source human and pose from reference human, as depicted in the top row of Fig. 1; ii) human novel view synthesis aims to synthesize new images of the human body, captured from different viewpoints, as illustrated in the middle row of Fig. 1; iii) the goal of appearance transfer is to generate a human image preserving the source face identity while wearing the clothes of the reference, as shown in the bottom row of Fig. 1 where each garment (upper-clothes or pants) might come from different people.

Taking human motion imitation as an example, existing methods can be roughly categorized into an image-to-image translation-based [8], [9], [10] pipeline and a warping-based pipeline [1], [2], [3], [11]. The image-to-image translation-based pipeline learns a person-specific mapping function from the human conditions, characterized by a skeleton, dense pose, and parsing result, to the image from a video with paired sequences of conditions and images. Thus, everybody needs to train their model from scratch, and a particular trained model cannot be applied to others. Besides, it is not accessible to be extended to other tasks, such as appearance transfer. To overcome this shortcoming, researchers have proposed

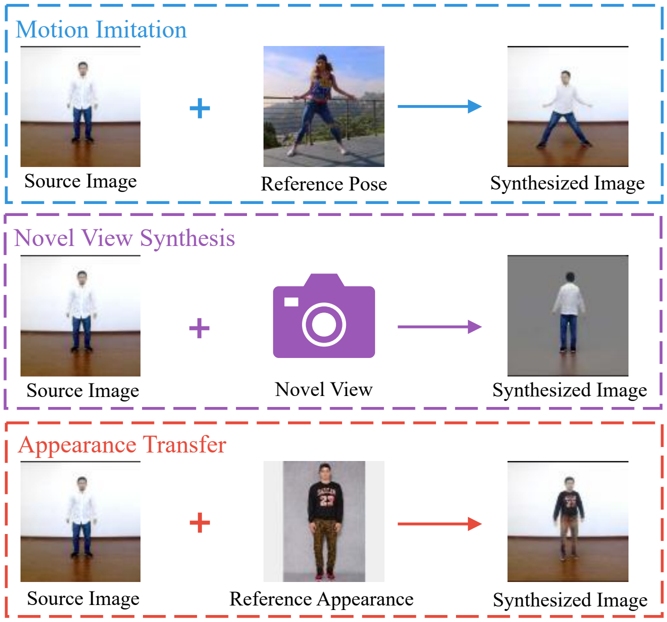
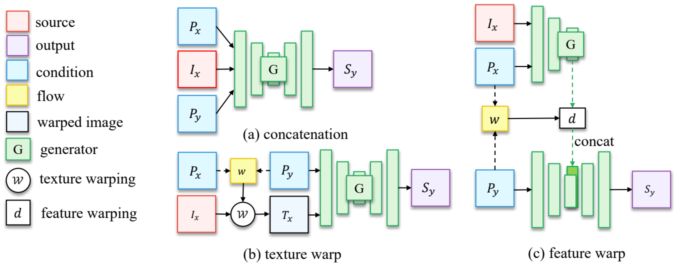


Fig. 1: Illustration of human motion imitation, novel view synthesis and appearance transfer. The 1^{st} row is the source image and the 2^{nd} row is reference condition, such as image or novel viewpoint of camera. The 3^{rd} row is the synthesized results.

- *Wen Liu is with School of Information Science and Technology, ShanghaiTech University, and Chinese Academy of Sciences, Shanghai Institute of Microsystem and Information Technology, and University of Chinese Academy of Sciences, China.*
- *Zhixin Piao, Zhi Tu, and Shenghua Gao are with School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China. Shenghua Gao is the corresponding author.*



(a) concatenation

(b) texture warp

(c) feature warp

Fig. 2: Three existing approaches to propagate the source information into the target condition. (a) early concatenation, concatenates the source image, the source condition, and the target condition into the color channel. (b) and (c) are texture and feature warping, respectively. The source image or its features are propagated into the target condition under a fitted transformation flow.

the warping-based methods, which warp the input images into the reference conditions (skeleton, dense pose, or parsing) and generate the desired image. So a trained model in these methods could be applied to other input images with different identities. We summarize the recent warping-based approaches in Fig. 2. An early work [2], shown in Fig. 2 (a), feeds the concatenated source image (with its pose condition) with the target pose condition into a network with an adversarial training to generate an image with the desired pose. However, direct concatenation does not consider the spatial layout, and it is ambiguous for the generator to place the pixel from a source image into the right position. Thus, it always results in a blurred image and loses the source identity. Later, inspired by the spatial transformer networks (STN) [12], a texture warping method [1], as shown in Fig. 2 (b), is proposed. It firstly fits a rough affine transformation matrix from the source and the reference key points, then uses an STN to warp the source image into the reference pose, and after that generates the final result based on the warped image. However, texture warping could not preserve the source information as well, in terms of the color, style, or face identity, because the generator might drop out the source information after several downsampling operations, such as stride convolution and pooling. Meanwhile, contemporary work [3], [11], [13] proposes to warp the deep features of the source images into the target poses rather than that in the image space, as shown in Fig 2 (c), named as feature warping. However, features extracted by an encoder in the feature warping cannot guarantee to characterize the source identity accurately, which consequently produces a blur or low-fidelity image inevitably.

The aforementioned existing methods encounter with challenges in generating realistic-looking images, due to three reasons: 1) diverse clothes in terms of texture, style, color, and high-structure face identity are difficult to be captured and preserved in their network architectures; 2) articulated and deformable human bodies result in a large spatial layout and geometric changes for arbitrary pose manipulations; 3) all these methods cannot handle multiple source inputs, such as in appearance transfer, different parts might come from different source people; 4) the generalization is not good when the inputs are out of the domain of training set because to synthesize photo-realistic images, all these methods apply the adversarial constraints of discriminators, which push the results similar to the distribution of training set.

In this paper, we follow the warping-based pipeline. To preserve the source details of the clothes and face identity, we propose a Liquid Warping Block (LWB) and an advanced version, Attentional Liquid Warping Block (AttLWB), to address the loss of the source information from three aspects: 1) a denoising convolutional auto-encoder is used to extract useful features that preserve the source information, including texture, color, style and face identity; 2) the source features of each local part are blended into a global feature stream by our proposed LWB and AttLWB, to preserve the source details further; 3) it supports multiple-source warping, such as in the appearance transfer that supports to warp the features of a head (local identity) from one source and that of a body from another, and aggregate them into a global feature stream; 4) a one/few-shot learning strategy is utilized to improve the generalization of the network.

In addition, existing approaches mainly rely on a 2D pose [1], [2], [3], a dense pose [14] and body a parsing result [11]. These methods only take care of the layout locations and ignore the personalized shape and limb (joints) rotations, which are even more essential than layout locations in human image synthesis. For example, in an extreme case that a tall man imitates the actions of a short person, if we the 2D skeleton, the dense pose and the body parsing condition will unavoidably change the height and the size of the tall one, as shown at the bottom of Fig. 9. To overcome these issues, we use a parametric statistical human body model, SMPL [15], [16], [17], [18], which disentangles a human body into the pose (joint rotations) and the shape. It outputs a 3D mesh (without clothes) rather than the layouts of joints and parts. Further, transformation flows can be easily calculated by matching the correspondences between two 3D triangulated meshes, which is more accurate and results in fewer misalignments than previous fitted affine matrix from keypoints [1], [3].

Based on the SMPL model and the Liquid Warping Block (LWB) or the Attentional Liquid Warping Block (AttLWB), our method can be further extended into other tasks, including human appearance transfer and novel view synthesis for free and one model can handle these three tasks. We summarize our contributions as follows: 1) we propose an LWB and an AttLWB to propagate and address the loss of the source information, such as texture, style, color, and face identity, in both the image and the feature space; 2) by taking advantages of both the LWB (AttLWB) and the 3D parametric model, our method is a unified framework for human motion imitation, appearance transfer, and novel view synthesis; 3) since the previous datasets [19], [20] have the limitation in the diversity of the poses, and can only be used for motion imitation, we build a dataset for these tasks, especially for human motion imitation in the video, and released all codes and datasets for further research convenience in the community.

This paper is an extension of our previous work [21]. We extend the framework in the following aspects:

i) our previous LWB [21] directly adds the warped multiple source features into the global features, and it will enlarge the magnitude of the features in the overlap area, thereby resulting in artifacts. To address this, motivated by the attention architecture [22], we propose a more advanced Attentional Liquid Warping Block (AttLWB). It firstly learns similarities of the global features among all multiple sources features, and then it fuses the multiple sources features by a linear combination of the learned similarities and the multiple sources in the feature spaces. Finally, to better propagate the source identity (style, color, and texture) into the global stream, we warp the fused source features to the global stream by the Spatially-Adaptive Normalization (SPADE) [23], which could further improve the final result;

ii) our previous network could not generalize well when the input images are far away from the training domain, as the interracial motion imitation. The reason might be that to generate images with high fidelity, an adversarial (GAN) loss is essential [1], [2], [3], [21], which pushes the generated images in the distribution of the training set. Considering that the input images are diverse in human races, face identities, and clothes styles, and it is infeasible to collect a dataset containing all these individuals. In the testing phase, once an individual is unique in face identity or clothes style, the well-trained network might produce a high-fidelity result similar to the training samples but does not preserve its own source identity in terms of face and clothes. To improve the generalization, inspired by the SinGAN [24] and the Few-Shot Adversarial Learning [25], we apply a one/few-shot adversarial learning to push the network to focus on the individual input with several steps of adaptation, namely personalization.

iii): our previous method successfully achieves decent results on 256×256 resolution, and in this version, based on the AttLWB and personalization, we could further achieve the high-fidelity results with a higher 512×512 and 1024×1024 resolution.

We organize the rest of this paper as follows: In Section 2, we summarize the related work of the Human Image Synthesis, including the motion imitation, the appearance transfer, and the novel view synthesis. In Section 3, we firstly introduce the essential modules of our proposed Attentional Liquid Warping GAN. The following are the training strategies, the loss functions, the one/few-shot personalization, and the inference details. In Section 4, extensive experiments on different datasets and tasks validate the effectiveness of our work. In Section 5, ablation studies and analysis are conducted to evaluate the impacts of different components. We conclude our work in Section 6.

2 Related work

2.1 Human Motion Imitation

We summarize the recent image-to-image translation-based and the warping-based methods as follows.

Image-to-Image translation-based methods. Esser *et al.* [26] use a Variational U-Net to learn a mapping function from a 2D skeleton to an image. Chan *et al.* [9] learn a mapping function from a 2D skeleton to an image by a pix2pixHD [27] with a specialized Face GAN and temporally coherent GAN. Wang *et al.* [28] propose a vid2vid framework and learn a mapping function from 2D dense pose to image. Meanwhile, Shysheya *et al.* [29] firstly build a full texture UV image of a person by multi-view cameras, then learn a mapping function from a 3D skeleton to part coordinates of the UV map and finally render a result based on the coordinates and the UV image. Contemporarily, Liu *et al.* [30] firstly use a monocular video to reconstruct a full 3D character model of a person with a static pose, then render the texture of each body parts and finally learn a mapping from synthetic to real images. However, all these methods train a mapping from keypoints or parts to each person's image and everybody needs to train their own model. This might limit its wide application.

Warping-based methods. Recent work is mainly based on the conditioned generative adversarial networks (CGAN) [1], [2], [14], [31], [32]. Their key technical idea is to combine the source image along with the source pose (2D skeleton) as inputs and generate a realistic image by GANs using a reference pose. The differences among those approaches are merely in network architectures, warping strategies, and adversarial losses. In [2], Ma *et al.* [2] directly concatenate the source image and the reference pose, and then design a U-Net [33] generator with a coarse-to-fine strategy to generate 256×256 images. Neverova *et al.* [14] replace the sparse 2D key points with the dense correspondences between the image and surface of the human body by the DensePose [34]. Si *et al.* [32] propose a multistage adversarial loss and separately generate the foreground (or different body parts) and background. Balakrishnan *et al.* [1] firstly fit an affine transformation matrix based on the source and the target 2D key points and then use a texture warping strategy to generate the foreground and the background separately. These work [3], [11], [13], [35], focus on the way of warping the source features into the target conditions, like skeleton or parsing. Besides, Li *et al.* [36] propose to learn a transformation flow from 2D key points and warp the deep features based on the learned transformations.

2.2 Human Appearance Transfer

Human appearance modeling or transfer is a vast topic, especially in the field of virtual try-on applications, from computer graphics pipelines [37] to learning based pipelines [4], [5]. Graphics based methods first estimate the detailed 3D human mesh with clothes via garments and 3D scanners [38] or multiple camera arrays [39], and then human appearance with clothes is capable of being conducted from one person to another based on the detailed 3D mesh. Although these methods can produce high-fidelity results, their cost, size, and controlled environment are unfriendly and inconvenient to customers. Recently, in the light of deep generative models, SwapNet [4] firstly learns a pose-guided clothing segmentation synthetic network, and then the clothing parsing results with texture features from the source image are fed into an encoder-decoder network to generate the image with the desired garment. In [5], the authors leverage a geometric 3D shape model combined with learning methods, swap the color of visible vertices of the triangulated mesh, and train a model to infer that of invisible vertices. Instead of estimating the 3D clothes by other sensors, in the MGN [40], the authors, train a network with 3D scans data and predict the body shape and clothing directly from 8 frames or a video. They apply the garment transfer based on the estimated 3D body mesh with clothes.

2.3 Human Novel View Synthesis

Novel view synthesis aims to synthesize new images of the same object or human body from arbitrary viewpoints. The core step of existing methods is to fit a correspondence map from the observable views to new views with convolutional neural networks. In [41], the authors use CNNs to predict appearance flow and synthesize new images of the same object by copying the pixel from a source image based on the appearance flow and they have achieved decent results of rigid objects like vehicles. The following work [42] proposes to infer the invisible textures based on appearance flow and adversarial generative network (GAN) [43], while Zhu *et al.* [7] argue that appearance flow-based method performs poorly on articulated and deformable objects, such as human bodies. They propose an *appearance-shape-flow* strategy to synthesize different views of human bodies – besides, Zhao *et al.* [6] design a GAN based method to synthesize high-resolution views in a coarse-to-fine way. Recently, in PiFu [44], the authors learn an implicit function with multi-layer perceptrons (MLPs) to digitize the human body and infer the 3D surfaces and texture from a single or multiple frames. The fully digitalized human body could synthesize a different view.

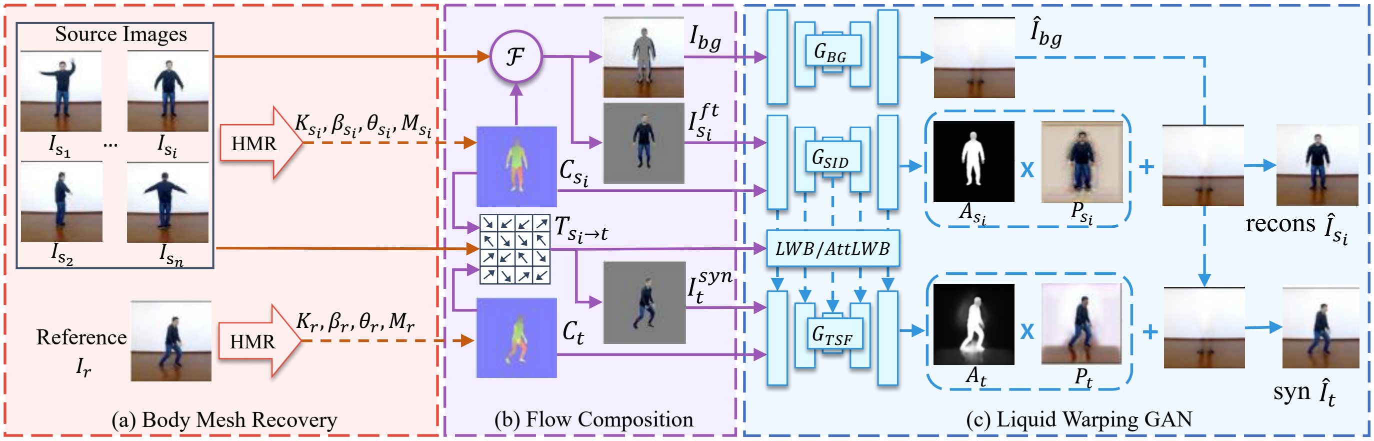


Fig. 3: The training pipeline of our method. We randomly sample a pair of images from a video, denoting the source and the reference image as I_{s_i} and I_r. **(a)** A body mesh recovery module will estimate the 3D mesh of each image and render their correspondence map, C_s and C_t; **(b)** The flow composition module will first calculate the transformation flow T based on two correspondence maps and their projected vertices in the image space. Then it will separate the source image I_{s_i} into a foreground image $I_{s_i}^{ft}$ and a masked background I_{bg}. Finally it warps the source image based on the transformation flow T and produces a warped image I_{syn}; **(c)** In the last GAN module, the generator consists of three streams, which separately generates the background image $\hat{I}_{bg}$ by G_{BG}, reconstructs the source image $\hat{I}_s$ by G_{SID} and synthesizes the target image $\hat{I}_t$ under the reference condition by G_{TSF}. To preserve the details of the source image, we propose a novel LWB and AttLWB (shown in Fig. 4) which propagates the source features of G_{SID} into G_{TSF} at several layers and preserve the source information, in terms of texture, style and color.

2.4 One/Few-shot Learning in Image Synthesize

Ding *et al.* [45] propose a generative adversarial one-shot face recognizer to synthesize new face images. Shaham *et al.* [24] introduce a SinGAN, an unconditional generative model from a single image. Zakharov *et al.* [25] apply the few-shot adversarial learning to generate the realistic talking head. In light of the success of the Meta-Learning in classification, reinforcement learning and network architecture search [46], [47], [48], Lee *et al.* [10] propose a MetaPix for the few-shot motion imitation. Wang *et al.* [49] extend the previous vid2vid [28] framework within a few-shot setting and make it capable of synthesizing videos of unseen subjects by leveraging few example images.

3 Our Approach

In this section, we first introduce the whole models of our framework. It contains three modules, a body mesh recovery, a flow composition, and a GAN module with the Liquid Warping Block (LWB) or the Attentional Liquid Warping Block (AttLWB). Then, the following are the training details and loss functions. Further, to improve the generalization, we introduce a one/few-shot learning strategy. We illustrate the details of how to apply our model to three tasks in the inference section (Sect. 3.6).

Once the model has been trained on one task, it can deal with other tasks as well. Here, we use motion imitation as an example, as shown in Fig. 3. Our framework supports multiple sources of inputs, denoting the source images as $\{I_{s_1}, I_{s_2}, ..., I_{s_n}\}$, and the reference image as I_r. Here, s_n is the number of source images. First, the body mesh recovery module will estimate the 3D mesh of I_{s_i} and I_r and render their correspondence maps, C_{s_i}, and C_t. Next, the flow composition module will calculate the transformation flow $T_{s_i \to t}$ of each source image to the reference, based on two correspondence maps and their projected mesh in image space. Each source image I_{s_i} is thereby decomposed as the foreground image $I_{s_i}^{ft}$ and the masked background $I_{s_i}^{bg}$. Since all source images share the same background, we randomly choose one of the masked backgrounds, denoted as I_{bg}. Simultaneously, each source image contributes its visible textures to warp a synthetic image I_t^{syn}, based on the transformation flow $T_{s_i \to t}$. The last (Attentional) Liquid Warping GAN module consists of three streams. It separately generates the background image by G_{BG}, reconstructs the source image $\hat{I}_{s_i}$ by G_{SID} and synthesizes the final result $\hat{I}_t$ under the reference condition by G_{TSF}. To preserve the details of source image, we propose the novel Liquid Warping Block (LWB) and Attentional Liquid Warping Block (AttLWB) which propagate the source features of G_{SID} into G_{TSF} at multiple layers.

3.1 Body Mesh Recovery Module

As shown in Fig. 3 (a), given the source image I_{s_i} and the reference image I_r, the role of this stage is to predict the kinematic pose (rotation of limbs) and shape parameters, as well as the 3D mesh of each image. In this paper, we use the HMR [17], [18] as the 3D pose and shape estimator due to its good trade-off between accuracy and efficiency. In HMR, an image is firstly encoded into a feature with $\mathbb{R}^{2048}$ by a ResNet-50 [50] and then followed by an iterative 3D regression network that predicts the pose $\theta \in \mathbb{R}^{72}$ and the shape $\beta \in \mathbb{R}^{10}$ of SMPL [16], as well as the weak-perspective camera $K \in \mathbb{R}^3$. SMPL is a 3D body model that can be defined as a differentiable function $M(\theta, \beta) \in \mathbb{R}^{N_v \times 3}$, and it parameterizes a triangulated mesh by $N_v = 6,890$ vertices and $N_f = 13,776$ faces with the parameters of a pose $\theta \in \mathbb{R}^{72}$ and a shape $\beta \in \mathbb{R}^{10}$. Here, the shape parameters β are the coefficients of a low-dimensional shape space learned from thousands of registered scans, and the pose parameters θ are the joint rotations that articulate the bones via forwarding kinematics. With such process, we will obtain the body reconstructive estimations of each source image, $\{K_{s_i}, \theta_{s_i}, \beta_{s_i}, M_{s_i}\}$ and those of reference image, $\{K_r, \theta_r, \beta_r, M_r\}$, respectively.

3.2 Flow Composition Module

Based on previous estimations, we first render a correspondence map and a weight index map for each source mesh M_{s_i} and the reference mesh M_r under the camera view of K_{s_i} and K_r. Here, we denote the source weight index map, the source and the target correspondence maps as W_{s_i}, C_{s_i} and C_t, respectively. In this paper, we use a fully differentiable renderer, Neural Mesh Renderer (NMR) [51]. We thereby project vertices of the source V_{s_i} into a 2D image space by a weak-perspective camera, $v_{s_i} = \pi(V_{s_i}, K_{s_i})$. Here, π is the weak-perspective projective function. Then, we calculate the barycentric coordinates of each mesh face and obtain $f_{s_i} \in \mathbb{R}^{N_f \times 2}$. Next, we calculate the transformation flow $T_{s_i \to t} \in \mathbb{R}^{H \times W \times 2}$ by matching the correspondences between the source correspondence map with its mesh face coordinates f_{s_i}. Here $H \times W$ is the size of the image. By the same means, we obtain the transformation flow $T_{r \to t}$ of the reference correspondence map. We describe the procedure to obtain the transformation flow in Algorithm 1. Consequently, a foreground image $I^{ft}_{s_i}$ and a masked background image $I^{bg}_{s_i}$ are derived from masking the source image I_{s_i} based on C_{s_i}. We randomly pick one of the masked backgrounds, denoted as I_{bg}, because all source images share the same background. Finally, we warp the visible textures of each source image I_{s_i} to the desired condition by the transformation flow $T_{s_i \to t}$ and thereby obtain a synthetic image I^{syn}_t, as depicted in Fig. 3.

Algorithm 1 The procedure of obtaining transformation $T_{s_i \to t}$.

Input: $W_{s_i}, V_{s_i}, F_{s_i}, K_{s_i}, C_{s_i}, C_t$.

- $K_{s_i} \in \mathbb{R}^{3 \times 1}$: source weak-perspective camera;
- $V_{s_i} \in \mathbb{R}^{N_v \times 3}$: N_v is the number of vertices;
- $F_{s_i} \in \mathbb{R}^{N_f \times 3}$: N_f is the number of faces;
- $W_{s_i} \in \mathbb{R}^{H \times W \times 3}$: the weight index map of source mesh, the value of each pixel indicates the barycentric weights of the triangulated faces in image space;
- $C_{s_i}(C_t) \in \mathbb{R}^{H \times W \times 1}$: the correspondence map of source and target mesh, and the value in each pixel indicates the face index of the mesh.

Output: $T_{s_i \to t} \in \mathbb{R}^{H \times W \times 2}$, the output transformation flow;

```
$v_{s_i} = \pi(V_{s_i}, K_{s_i})$ # projecting vertices of source $V_{s_i}$ into the 2D image space by the weak-perspective camera;
$tri_{s_i} = v_{s_i}[F_{s_i}] \in \mathbb{R}^{N_f \times 3 \times 2}$ # the triangulated faces with vertices in 2D image space;
$Vis_{s_i} \in \mathbb{R}^{N_f \times 1}$ # the face visibility;
for $f = 1$ to $N_f$ do
    $Vis_{s_i}(f) = 1$ if $f$ appears in $C_{s_i}$ else 0;
end for
initializing $T_{s_i \to t} \in \mathbb{R}^{H \times W \times 2}$;
for $i = 1$ to $H$ do
    for $j = 1$ to $W$ do
        $f = C_t(i, j)$ # the face index in current pixel;
        $T_{s_i \to t}(i, j) = W_{s_i}(i, j) \times tri_{s_i}(f)$, if $Vis_{s_i}(f)$ is 1;
    end for
end for
return $T_{s_i \to t}$.
```

3.3 Attentional Liquid Warping GAN

This stage synthesizes high-fidelity human images under the desired condition. More specifically, it 1) synthesizes the background image; 2) predicts the color of invisible parts based on the visible parts; 3) generates pixels of clothes, hairs, and others out of the reconstruction of SMPL.

Generator. Our generator works in a three-stream manner. One stream, named G_{BG}, works on the concatenation of the masked background image I_{bg} and the mask obtained by the binarization of C_{s_i} in the color channel to generate the realistic background image $\hat{I}_{bg}$, as shown in the top stream of Fig. 3 (c). The other two streams are the source identity stream, namely G_{SID} and the transfer stream, namely G_{TSF}. G_{SID} is a denoising convolutional auto-encoder that aims to guide the encoder to extract the features that are capable of preserving the source information. Together with the $\hat{I}_{bg}$, it takes the masked source foreground $I^{ft}_{s_i}$ and the correspondence map C_{s_i} as its inputs and reconstructs source foreground image $\hat{I}_s$. G_{TSF} stream synthesizes the final result, which receives the warped foreground by a bilinear sampler and the correspondence map C_t as its inputs. To preserve the source information, such as texture, style, and color, we propose a novel Liquid Warping Block (LWB), as well as its advanced version, Attentional Liquid Warping Block (AttLWB), that links the source with the target streams. They blend the source features from G_{SID} and fuses them into the transfer stream G_{TSF}, as shown at the bottom of Fig. 3 (c).

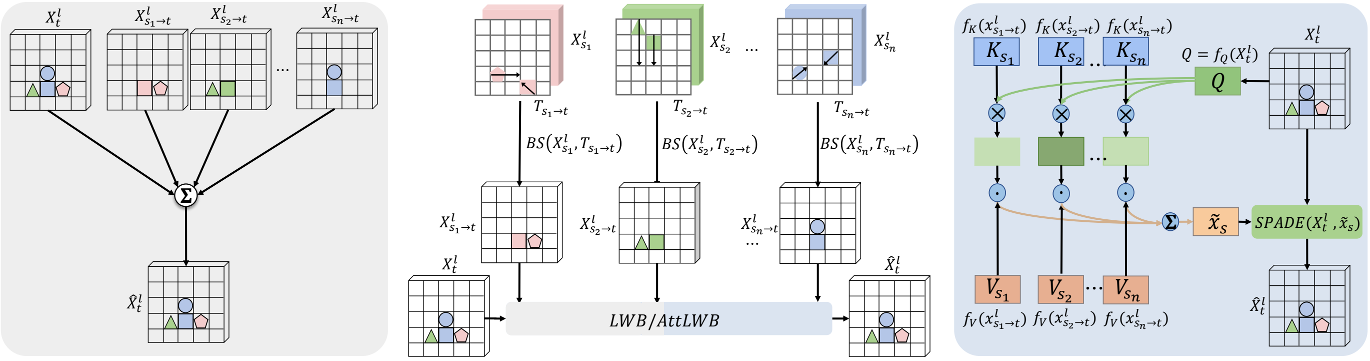
G_{BG} and G_{SID} have similar architectures with separate parameters and follow the structure of CycleGAN [52] with 6 residual blocks [53]. The details of kernel sizes and number of filters are illustrated in Fig. 5. G_{TSF} is a combination of a ResNet and a U-Net [33], named ResUnet. For G_{BG}, we directly regress the final background image, $\hat{I}_{bg}$, while for G_{SID} and G_{TSF}, we concretely generate an attention map A and a color map P, as shown in Fig. 5. The final image can be obtained as follows:

$$\begin{aligned} \hat{I}_{s_i} &= P_s \odot A_{s_i} + \hat{I}_{bg} \odot (1 - A_{s_i}) \\ \hat{I}_t &= P_t \odot A_t + \hat{I}_{bg} \odot (1 - A_t). \end{aligned} \tag{1}$$

Here, $\odot$ represents an element-wise multiplication. The total trainable parameters in the generator are $\theta_G = \{\theta_{BG}, \theta_{SID}, \theta_{TSF}, \theta_{AttLWB}\}$, with respect to G_{BG}, G_{SID}, G_{TSF} and AttLWB.

Discriminator. To push the discriminators to focus on different aspects of the generated images, such as the clothes on the human body and the face identity, we utilize a global-local content-orientation architecture. It consists of three sub-discriminators. The first one is a global discriminator, D_{Global}, which regularizes the entire generated $\hat{I}_t$ to be more realistic-looking. The rest two are a body discriminator D_{Body} and a face discriminator D_{Head}, and they push the cropped body area and the head (face) parts of the generated $\hat{I}_t$ to be realistic-looking. All of them are conditional discriminators, and they take the generated images and the correspondence map C_t as their inputs. We illustrate the details of our discriminators in Fig. 5. The total trainable parameters in the discriminators are $\theta^D = \{\phi_{Global}, \phi_{Body}, \phi_{Head}\}$.

Attentional Liquid Warping Block. One advantage of our proposed Liquid Warping Block (LWB) and Attentional Liquid Warping Block (AttLWB) is that it addresses the issue of multiple sources. For instance, in human motion imitation, the source images are multi-view inputs, and in the appearance transfer, different parts of garments come from different people. The different parts of features are aggregated into G_{TSF} by their transformation flow independently. As shown in Fig. 4, we denote $X^l_{s_1}$ and $X^l_{s_2}$ as the feature maps extracted by G_{SID} of different sources at the l^{th}


X_t^l
$X_{s_1\to t}^l$
$X_{s_2\to t}^l$
$X_{s_n\to t}^l$
Σ
$\hat{X}_t^l$
$X_{s_1}^l$
$X_{s_2}^l$
$X_{s_n}^l$
$T_{s_1\to t}$
$T_{s_2\to t}$
$T_{s_n\to t}$
$BS(X_{s_1}^l, T_{s_1\to t})$
$BS(X_{s_2}^l, T_{s_2\to t})$
$BS(X_{s_n}^l, T_{s_n\to t})$
LWB/AttLWB
$f_K(x_{s_1\to t}^l)$ $f_K(x_{s_2\to t}^l)$ $f_K(x_{s_n\to t}^l)$
K_{s_1} K_{s_2} K_{s_n}
$Q = f_Q(X_t^l)$
Q
$\tilde{x}_s$
$SPADE(X_t^l, \tilde{x}_s)$
V_{s_1} V_{s_2} V_{s_n}
$f_V(x_{s_1\to t}^l)$ $f_V(x_{s_2\to t}^l)$ $f_V(x_{s_n\to t}^l)$

(a) Add Warping Block (LWB)

(b) (Attentional) Liquid Warping Block

(c) Attentional Warping Block (AttLWB)

Fig. 4: Illustration of our LWB and AttLWB. They have the same structure illustrated in **(b)** but with separate AddWB (illustrated in **(a)**) or AttWB (illustrated in **(b)**). **(a)** is the structure of AddWB. Through AddWB, $\widehat{X}_t^l$ is obtained by aggregation of warped source features and features from G_{TSF}. **(b)** is the shared structure of (Attentional) Liquid Warping Block. $\{X_{s_1}^l, X_{s_2}^l, ..., X_{s_n}^l\}$ are the feature maps of different sources extracted by G_{SID} at the l^{th} layer. $\{T_{s_1\to t}, T_{s_2\to t}, ..., T_{s_n\to t}\}$ are the transformation flows from different sources to the target. X_t^l is the feature map of G_{TSF} at the l^{th} layer. **(c)** is the architecture of AttWB. Through AttWB, final output features $\widehat{X}_t^l$ is obtained with SPADE by denormalizing feature map from G_{TSF} with weighted combination of warped source features by a bilinear sampler (BS) with respect to corresponding flow $T_{s_i\to t}$.

layer and X_t^l is the feature map of G_{TSF} at the l^{th} layer. Each part of the source feature is warped by their transformation flow and aggregated into the features of G_{TSF}. We use a bilinear sampler (BS) to warp the source features $X_{s_1}^l$ and $X_{s_2}^l$ with respect to corresponding transformation flows, $T_{s_1\to t}$ and $T_{s_2\to t}$. The way to aggregate the warped source features into the global stream is the main difference between LWB and AttLWB.

LWB, as illustrated in Fig. 4 (a), directly uses an element-wise addition among all features and the fuses the global features as:

$$\begin{aligned} X_{s_i\to t}^l &= BS(X_{s_i}^l, T_i) \\ \widehat{X}_t^l &= \sum_{i=1}^{s_n} X_{s_i\to t}^l + X_t^l. \end{aligned} \tag{2}$$

However, LWB will enlarge the magnitude of the features in the overlap area, and thereby result in artifacts. To address this, motivated by the attention architecture [22], we propose a more advanced Attentional Liquid Warping Block (AttLWB), as shown in Fig. 4 (c). It firstly learns similarities of the global features among all multiple source features, and then it fuses the multiple source features by the linear combination of the learned similarities and the multiple sources in feature space. Finally, to better propagate the source identity (style, color, and texture) into the global stream, we use the SPADE [23] to denormalize the feature map of G_{TSF} with the fused source features to obtain the global stream, which could further improve the final result. We describe the entire procedures of AttLWB in Algorithm 2.

Algorithm 2 The procedure of our AttLWB.

Input: $\{T_{s_1\to t}, ..., T_{s_n\to t}\}$, $\{X_{s_1}^l, ..., X_{s_n}^l\}$, and X_t^l.

- $\{T_{s_1\to t}, ..., T_{s_n\to t}\}$: the transformation flows from different sources to the target;
- $\{X_{s_1}^l, ..., X_{s_n}^l\}$: the feature maps extracted by G_{SID} of different sources at the l^{th} layer;
- X_t^l: the feature map of G_{TSF} at the l^{th} layer;

Output: $\widehat{X}_t^l$, the output features;

```
$X_{s_i\to t}^l = BS(X_{s_i}^l, T_{s_i\to t})$ # warping each source feature;
$Q = f_Q(X_t^l)$ # query embeddings;
$K = [f_K(X_{s_1\to t}^l), ..., f_K(X_{s_n\to t}^l)]$ # key embeddings;
$V = [f_V(X_{s_1\to t}^l), ..., f_V(X_{s_n\to t}^l)]$ # value embeddings;
$\tilde{x}_s = Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V$ # fused source features, $d_k$ is the number of channels of $K$;
$\widehat{X}_t^l = SPADE(X_t^l, \tilde{x}_s)$ # conditioned on $\tilde{x}_s$;
return $\widehat{X}_t^l$;
```

3.4 Training Details and Loss Functions

In this part, we will introduce the loss functions and how to train the whole system. For the body recovery module, we follow the network architecture and loss functions of HMR [17], [18]. Here, we use a pre-trained (off-the-shelf) SMPL estimator.

Note that our proposed Attentional Liquid Warping GAN is a unified framework for motion imitation, appearance transfer, and novel view synthesis. Therefore once we have trained the model on one task, it is capable of being applied to other tasks. These three tasks share the same training pipeline in our method, except for the way to sample the source the reference images. In motion imitation, we randomly sample $s_n + 1$ images from each video with difference poses and set the first s_n ones as the source images $\{I_{s_1}, ..., I_{s_n}\}$ and the other one as the reference I_r. In appearance transfer, we need to sample $s_n + 1$ images with the same person identity wearing different clothes, while in novel view synthesis, we need to sample $s_n + 1$ images of the same person under the different camera of views. In our experiments, we train a model for motion imitation and then apply it to appearance transfer and novel view synthesis.

The whole loss function of the generator contains four terms, which are perceptual loss [54], face identity loss, attention regularization loss, and adversarial loss.

Perceptual Loss. It regularizes the reconstructed source image $\hat{I}_{s_i}$ to the ground truth I_{s_i} and pushes the generated target image $\hat{I}_t$ and the reference image I_r to be closer in a VGG [55] feature

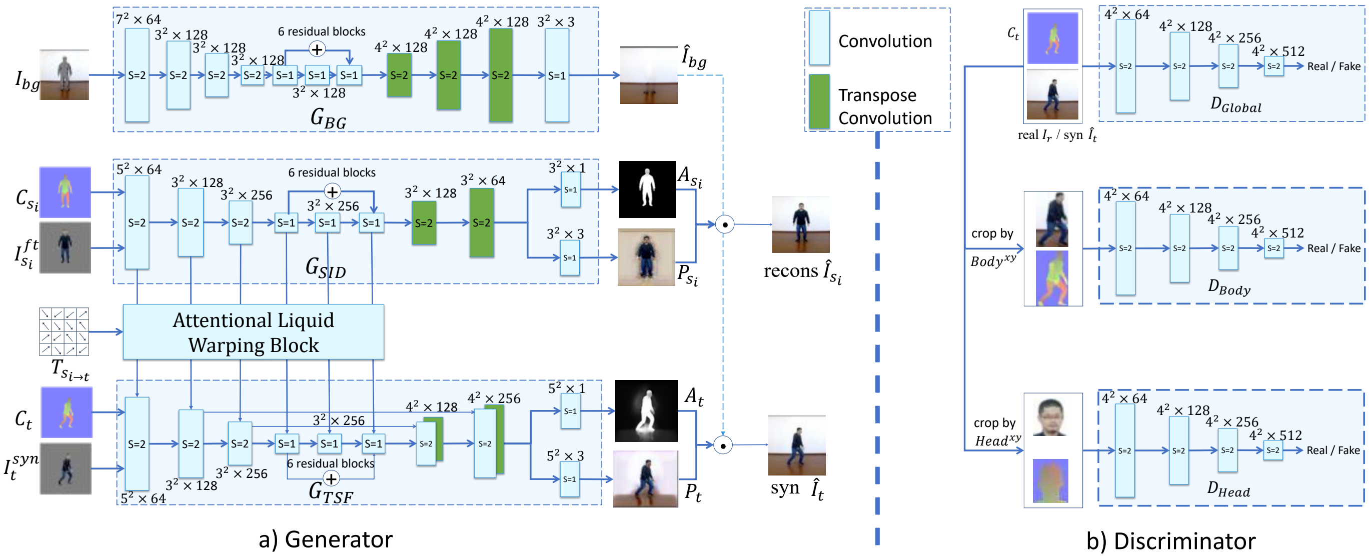


Fig. 5: The details of network architectures of our Attentional Liquid Warping GAN, including the generator and the discriminator. Here s represents the stride size in convolution and transposed convolution.

subspace. Its formulation is given as follows:

$$L_p = \frac{1}{s_n}\sum_{i=1}^{s_n} \|\hat{I}_{s_i} - I_{s_i}\|_1 + \|f(\hat{I}_t) - f(I_r)\|_1. \qquad (3)$$

Here, f is a pre-trained VGG-19 [55] on ImageNet [56].

Face Identity Loss. It regularizes the cropped face from the synthesized target image $\hat{I}_t$ to be similar to that from the image of ground truth I_r, which pushes the generator to preserve the face identity. It is shown as follows:

$$L_f = \|g(\hat{I}_t) - g(I_r)\|_1. \qquad (4)$$

Here, g is a pre-trained SphereFaceNet [57].

Adversarial Loss. It pushes the distribution of synthesized images to the distribution of real images. We use a $LSGAN_{-110}$ [58] loss in a way like PatchGAN over all discriminators, D_{Global}, D_{Body} and D_{Head}. They push the entire generated images, cropped body area, and head (face) parts to be realistic-looking. We denote the bounding box of head and body as $head^{xy}$ and $body^{xy}$ in the ground-truth I_r, respectively, and we calculate them by the projected vertices in the image space. $\hat{I}_t^b$, I_r^b and C_t^b are the cropped bodies from the generated image, the reference image and the correspondence map, based on bounding box of body, $body^{xy}$. $\hat{I}_t^h$, I_r^h and C_t^h are the corresponding cropped heads with respect to the bounding box of head, $head^{xy}$. We arrive at the total adversarial loss as follows:

$$\begin{aligned} L_{adv}^G = &\sum D_{Global}(\hat{I}_t, C_t)^2 + \sum D_{Body}(\hat{I}_t^b, , C_t^b)^2 \\ &+ \sum D_{Head}(\hat{I}_t^h, C_t^h)^2 \end{aligned} \qquad (5)$$

Attention Regularization Loss. It regularizes the attention map A_t and A_{s_i} to be smooth and prevents them from saturating. Considering that there is no ground truth of attention map A or color map P, they are learned from the resulting gradients of above losses. However, the attention masks can easily saturate to 1 which prevents the generator from working. To alleviate this situation, we regularize the mask to be closer to the silhouettes S rendered from a 3D body mesh. Since the silhouettes is a rough map and it contains the body mask without clothes and hair, we addtionaly introduce a Total Variation Regularization [59] over A to compensate the shortcomings of silhouettes. It is shown as:

$$\begin{aligned} L_a &= \|A_s - S_s\|_2^2 + \|A_t - S_t\|_2^2 + TV(A_s) + TV(A_t) \\ TV(A) &= \sum_{i,j}[A(i,j) - A(i-1,j)]^2 + [A(i,j) - A(i,j-1)]^2. \end{aligned} \qquad (6)$$

For the generator, the full objective function is shown as follows, and λ_p, λ_f and λ_a are the weights of perceptual, face identity and attention losses, respectively.

$$L^G = \lambda_p L_p + \lambda_f L_f + \lambda_a L_a + L_{adv}^G. \qquad (7)$$

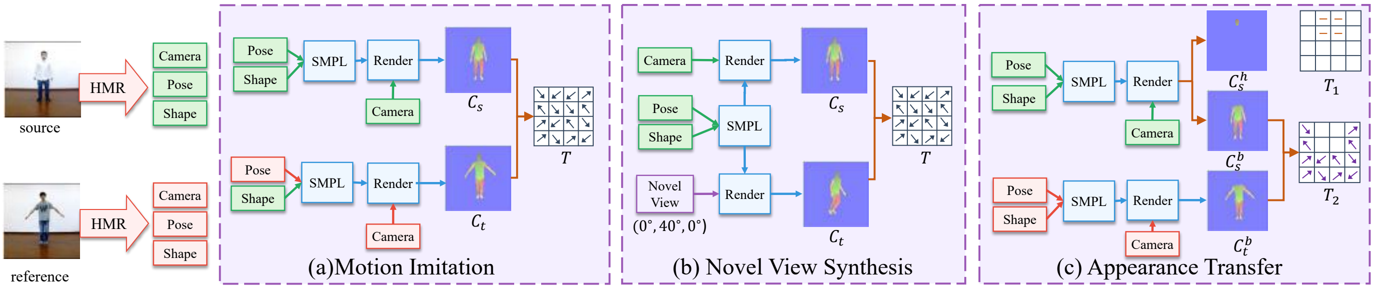
For discriminator, the full objective function is

$$\begin{aligned} L^D = &\sum[D_{Global}(\hat{I}_t, C_t) + 1]^2 + \sum[D_{Global}(I_r, C_t) - 1]^2 \\ &+ \sum[D_{Body}(\hat{I}_t^b), C_t^b) + 1]^2 + \sum[D_{Body}(I_r^b, C_t^b) - 1]^2 \\ &+ \sum[D_{Head}(\hat{I}_t^h, C_t^h) + 1]^2 + \sum[D_{Head}(I_r^h, C_t^h) - 1]^2. \end{aligned} \qquad (8)$$

3.5 One/Few-shot Personalization by Fine-tunning

Though we can train our model on a large dataset, to a certain degree, with diverse people and clothes, however, such a generator is still hard to be well-generalized to the inputs out of the domain of training set. After all, it is infeasible to build a universal dataset and generator to handle the diverse face identities, styles of clothes, and backgrounds. To improve the generalization, inspired by the SinGAN [24] and the Meta-learning [10], [25], [46], [49], we apply the one/few-shot adversarial learning to push the network to focus on each individual by several steps of fast personal adaptation. In real application scenarios, the user might only provide a little number (s_n) of their photos with different views or poses, and in an extreme case, there is only one image accessible. In this paper, we focus on the setting where there are no more than eight images ($s_n \leq 8$) [25] available in the testing phase.

Specifically, we first train our model, including a generator and a discriminator, on a combined large dataset, and consequently obtain the generator's pre-trained parameters, θ_G^M, and the discriminator's pre-trained parameters, θ_D^M. Then, for each



(a)Motion Imitation (b) Novel View Synthesis (c) Appearance Transfer

Fig. 6: Illustration of calculating the transformation flows of different tasks during the testing phase. The left is the disentangled body parameters by the Body Recovery module of both source and reference images. The right is the different implementations to calculate the transformation flow in different tasks.

specific person P_i with s_n images, we learn the person-specific generator $\theta_G^{P_i}$ and discriminator $\theta_D^{P_i}$ from the s_n images by fine-tuning the pre-trained model. This process is called one/few-shot personalization. To further push the generator from the pre-trained θ_G^M to the person-specific $\theta_G^{P_i}$, we discard the pre-trained parameters of the discriminator θ_D^M, and we train the person-specific discriminator $\theta_D^{P_i}$ from scratch. The overall loss functions in the personalization phase are similar to that in the training phase, except for the adversarial loss. Since there are only a few images ($s_n \leq 8$), to avoid overfitting and reduce the time consumption of each iteration in personalization, we only use the global discriminator.

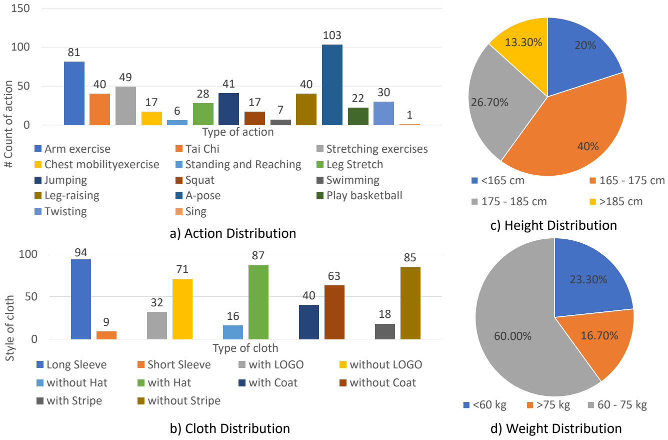
3.6 Inference

After we conduct personalization, the person-specific generator can be applied to all three tasks. The difference lies in the transformation flow computation, due to the different conditions of various tasks. The remaining modules, Body Mesh Recovery and Liquid Warping GAN (Attentional Liquid Warping GAN) are all the same. The followings are the details of each task of the Flow Composition module in the testing phase.

Motion Imitation. We firstly copy the value of pose parameters of the reference θ_r into that of the source and get the synthetic parameters of SMPL, as well as the 3D mesh, $M_t = M(\theta_r, \beta_s)$. Next, we render a correspondence map of the source mesh M_s and that of the synthetic mesh M_t under a camera view K_s. Here, we denote the source and the synthetic correspondence map as C_s and C_t, respectively. Then, we project the source vertices into the 2D image space by a weak-perspective camera, $v_s = \pi(V_s, K_s)$. Here, π is the weak-perspective projective function. Next, we calculate the barycentric coordinates of each mesh face and have $f_s \in \mathbb{R}^{N_f \times 2}$. Finally, we calculate the transformation flow $T \in \mathbb{R}^{H \times W \times 2}$ by matching the correspondences between the source correspondence map with its mesh face coordinates f_s and the synthetic correspondence map. It is shown in Fig. 6 (a).

Novel View Synthesis. Given a new camera view, in terms of a rotation R and a translation t. We firstly calculate the 3D mesh under the novel view, $M_t = M_s R + t$. The consequential operations are similar to that of motion imitation. We render a correspondence map of the source mesh M_s and that of the novel mesh M_t under a weak-perspective camera K_s and calculate the transformation flow $T \in \mathbb{R}^{H \times W \times 2}$, as depicted in Fig. 6 (b).

Appearance Transfer. We need to “copy” the clothes on the body from the reference image while keeping the head (face, eye, hair and so on) identity of the source. We split the transformation flow T into two sub-transformation flows, source flow T_1 and referent flow T_2. We denote the head mesh as $M^h = (V^h, F^h)$ and the body mesh as $M^b = (V^b, F^b)$. Here, $M = M^h \cup M^b$. For T_1, We firstly project the head mesh M_s^h of source into the image space and thereby obtain the silhouettes, S_s^h. Then, we create a mesh grid, $G \in \mathbb{R}^{H \times W \times 2}$. Next, we mask G by S^h and derive $T_1 = G \odot S^h$. Here, $\odot$ represents an element-wise multiplication. For T_2, it is similar to that in motion imitation. We render the correspondence map of the source body M_s^b and that of the reference M_t^b, denoted as C_s^b and C_t^b, respectively. Finally, we calculate the transformation flow T_2 based on the correspondences between C_s^b and C_t^b. We illustrate it in Fig. 6 (c).



a) Action Distribution c) Height Distribution b) Cloth Distribution d) Weight Distribution

Fig. 7: The statistic information of iPER dataset, including the action, clothes, height and weight distribution of the actors.

4 Experiments

4.1 Dataset

iPER. To evaluate the performance of our proposed method of motion imitation, appearance transfer, and novel view synthesis, we build a new dataset with diverse styles of clothes in videos, named Impersonator (iPER) dataset. There are 30 subjects of different conditions of shape, height, and gender. Each subject wears different clothes and performs an A-pose video and a video with random actions. There are 103 clothes in total. The whole dataset contains 206 video sequences with 241,564 frames. We split it into training/testing set at a ratio of 8:2 according to the different clothes. All the clothes and 29% of the actors in the testing set do not appear in the training set. We illustrate the details of the iPER dataset in classes of actions, styles of clothes, weight, and height distributions of actors in Fig. 7. We show some samples in the first two rows of Fig. 8.

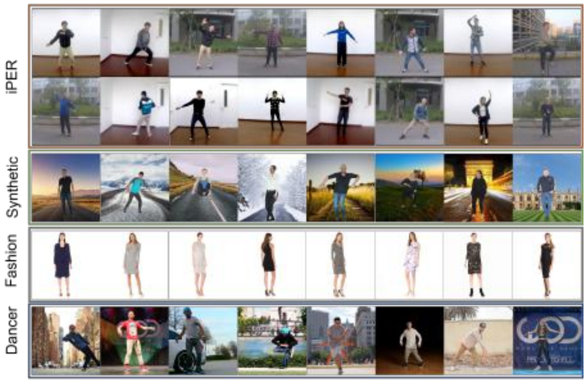


Fig. 8: The samples of four datasets. The first two rows are the samples from iPER dataset. The third row is the samples from the MotionSynthetic dataset and the fourth row is that from FashionVideo dataset. The last row is the samples from Youtube-Dancer-18 dataset.

MotionSynthetic. We also make up a synthetic dataset, named MotionSynthetic, for the convenience of evaluation, especially for human appearance transfer and novel view synthesis, because we can synthesize the ground truth images with different views and wearing garments by the modification of meshes. This dataset borrows 24 human meshes from people snapshot [60] and 96 human meshes from MultiGarments [40]; thus, 120 meshes in total. All of these meshes with UV texture images have been registered in SMPL [16]. For each mesh, we choose a pose sequence from Mixamo and a background image from the Internet. Based on these materials (mesh, UV image, pose sequence, and background image), we render the synthetic images by NMR [51], resulting in 39,529 frames in total. We split it into training/testing set at a ratio of 8:2 according to the different meshes and illustrate some synthetic images in the 3rd rows of Fig. 8.

FashionVideo. It contains 500 training and 100 testing videos with a single female model wearing fashionable clothes [20]. Each video has around 350 frames. The clothes and textures are diverse, while there are few types of gestures, with only a few standard poses for the models. Also, this dataset lacks diversity in background, and all the backgrounds are black. We display some samples in the 4th row of Fig 8.

Youtube-Dancer-18. To further validate the effectiveness and generalization of our method, we evaluate our method on the in-the-wild internet videos, Youtube-Dancer-18 [10]. It consists of 18 videos, with people dancing, downloaded from Youtube, and each of them lasts from 4 to 12 minutes. We follow the setting with MetaPix [10] that we sample frames with 30 FPS and only use $s_n \leq 8$ frames from training sequences for personalization and then apply the evaluation on the testing sequences. Some samples are shown at the bottom of Fig. 8. It needs to be mentioned that we do not train the model in this dataset. We only sample s_n frames for personalization and directly test on this dataset to evaluate the generalization over all methods.

4.2 Implementation Details

We train our Attentional Liquid Warping GAN on a combined dataset consisting of the iPER, MotionSynthetic, and FashionVideo dataset and perform evaluations among these three datasets. To evaluate our methods' generalization, we also perform tests on an additional Youtube-Dancer-18 dataset without training on it. We crop all images based on the bounding box of the human body, rescale the cropped images with keeping the original ratio of height and width, and then pad them into a 512×512 resolution. We normalize the color space of all images to [-1, 1]. In our experiments, including the training and personalization phase, we use the Adam [61] based Stochastic Gradient Descent optimizer for both generators and discriminators. λ_p, λ_f and λ_a are 10.0, 5.0 and 2.5, respectively.

i): In the training phase, we randomly sample $s_n + 1$ images from each video and set the first s_n ones as the source images $\{I_{s_1}, ..., I_{s_n}\}$, and the other one as the reference I_r. We fix $s_n = 2$ and the mini-batch size to be 2. There are two training epochs. We fix the first quarter training session with a learning rate as 0.0001 and gradually decrease it to 0.00001 in the end.

ii): In the personalization and testing phase, s_n could be flexible, and because of the memory limitation of the GPU devices, in our experiments, we set $s_n \in \{1, 2, 4, 8\}$. Besides, I_r lies in the set of source images $\{I_{s_1}, ..., I_{s_n}\}$. We fix the learning rate as 0.0001 and take $T = 100$ steps for personalization.

4.3 Results of Human Motion Imitation

Evaluation Metrics. We propose an evaluation protocol of the testing set of the iPER, MotionSynthetic, FashionVideo, and Youtube-Dancer-18 datasets, and it can indicate the performance of different methods in terms of different aspects. The details are listed in followings:

1): In each video with actor P_i, $\{I_1^{P_i}, ..., I_t^{P_i}, ..., I_L^{P_i}\}$, we select eight images as candidate images with different views, such as frontal, sideways or back. Here, L is the number of frames.

2): We choose $s_n \leq 8$ images as sources, $\{I_{s_1}^{P_i}, ..., I_{s_n}^{P_i}\}$, from the eight candidate images for personalization. For a fair comparison with other methods [1], [2], [3], [35], [36], which only use a single source image, we separately report the results on $s_n = 1$ (one-shot setting) and $2 \leq s_n \leq 8$(few-shot setting).

3): After personalization, we perform self-imitation that each actor P_i imitates actions from images of themselves, with $I_t^{P_i}$ as the reference image. We denote $\hat{I}_t^{P_i \rightarrow P_i}$ as the synthesized image referring to $I_t^{P_i}$. As for criterion, we use PSNR, SSIM [62], Learned Perceptual Similarity (LPIPS) [63], Body-CS and Face-CS to measure the similarities between $\hat{I}_t^{P_i \rightarrow P_i}$ and $I_t^{P_i}$.

Body-Cosine-Similarity (Body-CS): is the distance between the cropped person region of the synthesized image and that of the ground-truth image. In particular, it firstly uses a YOLOv3 [64] detector to get the person bounding box of the synthesized and ground-truth image. Then, we crop the person patches according to the bounding boxes. Finally, we use a pre-trained Person re-identification (ReID) model, OS-Net [65], to get the embedding features of the cropped person patches, and then we normalize the features and calculate the cosine similarity between the features to acquire the Body-CS.

Face-Cosine-Similarity (Face-CS): similar to Body-CS, it is the distance between the cropped face region of the synthesized image and that of the ground-truth image. Specifically, we firstly use an MTCNN [66] face detector to get the face bounding boxes

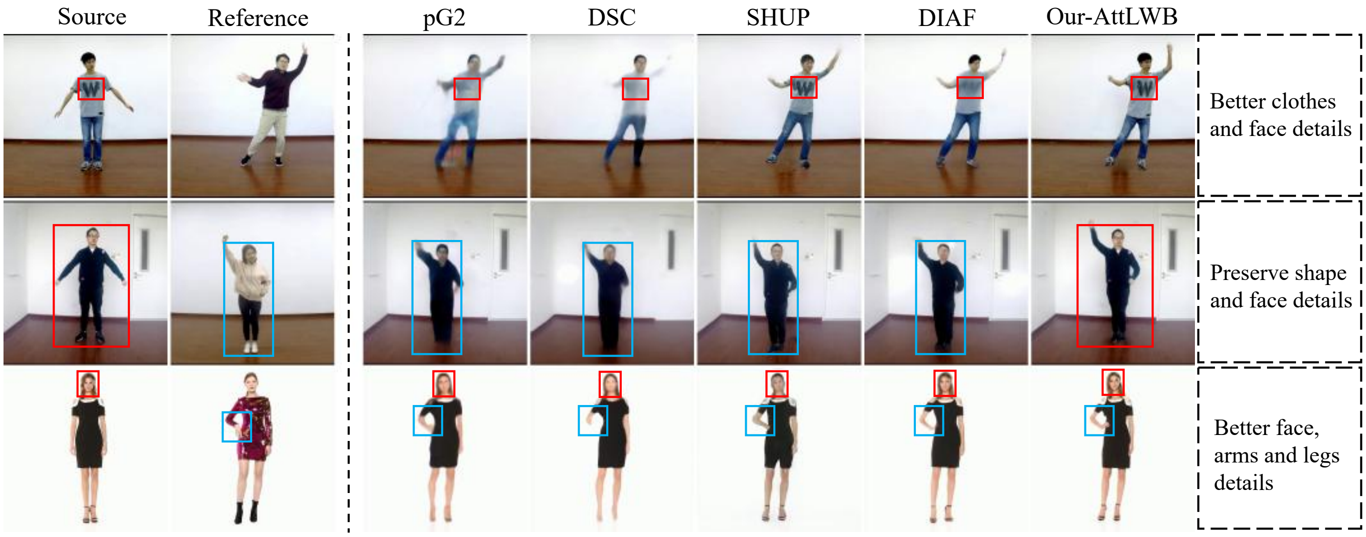

Source
Reference
pG2
DSC
SHUP
DIAF
Our-AttLWB
Better clothes and face details
Preserve shape and face details
Better face, arms and legs details

Fig. 9: Comparison of our method with others of motion imitation on the iPER and FashionVideo dataset (zoom-in for the best of view). All results are in 512 × 512 resolution. 2D pose-guided methods pG2 [2], DSC [3] SHUP [1] and DIAF cannot preserve the clothes details, face identity and shape consistency of source images. We highlight the details by red and blue rectangles.

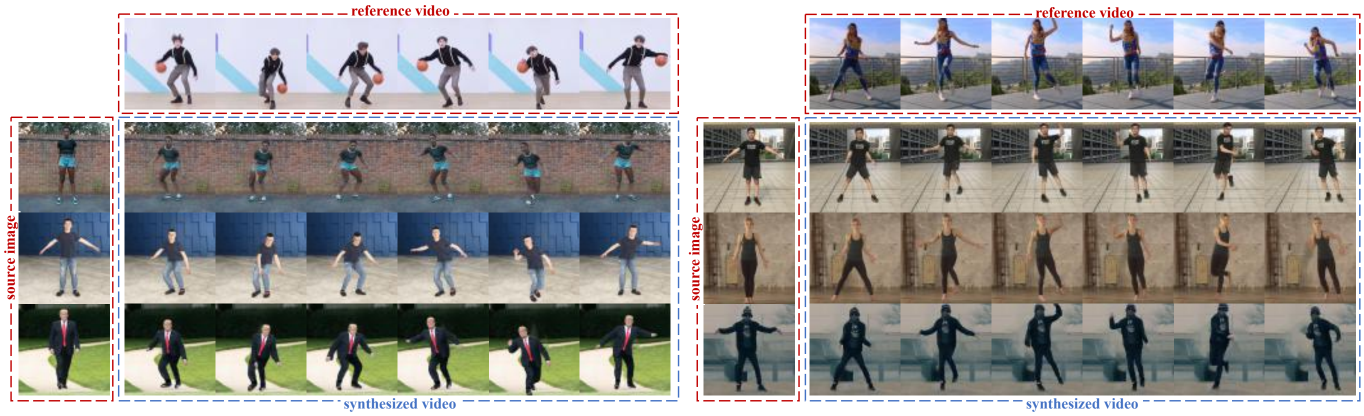


Fig. 10: Examples of motion imitation from our proposed methods (zoom-in for the best of view). All results are in 512 × 512 resolution. Our method could produce high-fidelity images that preserve the face identity, shape consistency and clothes details of source. We recommend accessing the supplementary material for more results in videos.

of the synthesized and ground-truth images. Then, we crop the face regions according to the bounding boxes. Finally, we uses a pre-trained face recognition model [67], to get the embedding features of the cropped face patches, and then we normalize the features and calculate the cosine similarity between the normalized features to obtain the Face-CS.

4): We also conduct cross-imitation that an actor P_i imitates actions from others, such as P_j. We denote $\{\hat{I}_1^{P_i \to P_j}, ..., \hat{I}_L^{P_i \to P_j}\}$ as a sequence of synthesized images referring to $\{I_1^{P_j}, ..., I_L^{P_j}\}$ and $\{I_{s_1}^{P_i}, ..., I_{s_n}^{P_i}\}$ as the sequence of real images. Since there is no ground-truth of synthesized images for the similarities metrics as mentioned above, here, we use a Fréchet Inception Distance (FID) [68] to measure perceptual realism. It calculates the distance between the set of synthesized images and that of real images. We further propose the Fréchet Distance of a pre-trained ReID model, OS-Net [65], namely Body-FD and that of a face recognition model, namely Face-FD. We also collect L consecutive frames from the actor P^i, denoted as $\{I_1^{P_i}, ..., I_L^{P_i}\}$, then calculate the Body-CS and Face-CS as aforementioned.

Quantitative Comparison with Other Methods under One-shot Setting. We compare the performance of our method with that of existing methods, including PG2 [2], SHUP [1], DSC [3], DIAF [36] and PATB [35]. We train all these methods on a combined dataset with the iPER, MotionSynthetic, and FashionVideo dataset and apply the evaluation protocol with the one-shot setting mentioned above to these methods. We report the results in Table 1, and our method outperforms others on all the metrics except SSIM, for which a higher numerical value does not necessarily mean a better quality of an image as reported in [63].

Quantitative Comparison with Other Methods under Few-shot Setting. We compare the performance of our method with pix2pixHD [27], SPADE [23], MetaPix pix2pixHD and MetaPix SHUP [10] under this setting. Here, we report the results on the Youtube-Dancer-18 dataset with the number of source images s_n being 2 in Table 2 and our method outperforms others.

Qualitative Comparison. Besides, we also analyze the generated images and make comparisons between ours and the above methods. From Fig. 9, we find that 1) the above methods that

TABLE 1: **One-shot** average results for human motion imitation of different methods on the iPER, MotionSynthetic and FashionVidieo dataset. ↑ means the larger the better, and ↓ is on the contrary. A higher SSIM may not mean a better quality of an image [63].

	Self-Imitation					Cross-Imitation				
	PSNR↑	SSIM↑	LPIPS↓	Body-CS↑	Face-CS↑	Face-CS↑	Face-FD↓	Body-CS↑	Body-FD↓	FID↓
PG2 [2]	23.699	0.876	0.130	0.744	0.085	0.148	429.142	0.709	240.429	119.378
SHUP [1]	23.979	**0.881**	0.080	0.855	0.288	0.297	243.599	0.820	80.973	51.823
DSC [3]	20.782	0.732	0.331	0.695	0.139	0.204	407.070	0.673	273.103	150.082
DIAF [36]	22.753	0.829	0.108	0.851	0.390	0.364	166.560	0.808	102.807	63.528
PATB [35]	20.387	0.798	0.169	0.738	0.129	0.363	218.333	0.731	259.135	136.911
Our-LWB	23.932	0.843	0.089	0.901	0.560	0.538	99.258	0.862	48.619	32.370
Our-AttLWB	**24.513**	0.856	**0.074**	**0.911**	**0.591**	**0.564**	**73.217**	**0.869**	**44.022**	**30.503**

TABLE 2: **Few-shot** results for human motion imitation of different methods on the Youtube-Dancer-18 dataset. The number of source images s_n is 2. ↑ means the larger the better, and ↓ represents the smaller the better.

	Self-Imitation					Cross-Imitation				
	PSNR↑	SSIM↑	LPIPS↓	Body-CS↑	Face-CS↑	Face-CS↑	Face-FD↓	Body-CS↑	Body-FD↓	FID↓
pix2pixHD [27]	11.134	0.196	0.633	0.616	0.106	0.136	221.661	0.565	266.552	175.574
SPADE [23]	8.984	0.120	0.780	0.535	0.106	0.131	294.672	0.513	431.670	304.698
MetaPix Pix2PixHD [10]	14.052	0.385	0.550	0.549	0.134	0.187	277.555	0.523	441.495	257.457
MetaPix SHUP [10]	18.857	**0.649**	0.269	0.765	0.234	0.191	185.363	0.693	160.485	83.501
Our-LWB	19.485	0.642	0.245	0.830	0.413	0.355	96.280	0.738	102.075	70.743
Our-AttLWB	**19.691**	**0.649**	**0.232**	**0.831**	**0.437**	**0.380**	**82.053**	**0.743**	**99.575**	**65.454**

use 2D pose-guided inputs change the body shape of the source. For example, in the 2^{nd} row of Fig. 9, the scenario is a tall person imitating motion from a short person, and baseline methods change the height of the source body. However, our method is capable of keeping the body shape unchanged because our method disentangles the pose and the personalized shape of each actor. 2) In the light of our proposed AttLWB (LWB) and face identity loss, our method is more powerful in terms of preserving source identities, such as the face identity and cloth details of source than other methods, as shown in the 1^{st} and 2^{nd} row of Fig. 9. 3) Our method also produces high-fidelity images in the cross-imitation setting (imitating actions from others), which we illustrate in Fig. 10. As we can see in Fig. 10, the face identity, and clothes details, in terms of texture color and style, are preserved well. It shows that our method can achieve decent results in cross imitation even when the reference image comes from the Internet, which is out of the domain of our training dataset.

4.4 Results of Human Appearance Transfer

It is worth emphasizing that once the model has been trained, it can directly be applied in three tasks, including motion imitation, appearance transfer, and novel view synthesis. We conduct the experiments on the iPER dataset.

Evaluation Metrics. In the iPER dataset, subjects might wear different clothes, and we sample the same person's images with different clothes as the source and the reference image. We use aforementioned PSNR, SSIM [62], LPIPS [63], Body-CS and Face-CS as the metrics.

Quantitative Results. We report the results of our methods with LWB and AttWLB on the iPER dataset in Table 3. The results show that Attentional Liquid Warping Block (AttLWB) is slightly better than the LWB.

Qualitative Results. We randomly pick some examples displayed in Fig. 11. The face identity and clothes details, in terms of texture, color, and style, are preserved well by our method. It demonstrates that our method can achieve decent results in appearance transfer, even when the reference image comes from the Internet and is out of the domain of the iPER dataset, such as the last five columns in Fig. 11.

TABLE 3: Results for human appearance transfer of our LWB and AttLWB, on the iPER dataset. Here, we report the PSNR, SSIM, LPIPS, Body-CS and Face-CS. ↑ means the larger the better. A higher SSIM may not mean a better quality of an image [63].

	PSRN↑	SSIM↑	LPIPS↓	Body-CS↑	Face-CS↑
Our-LWB	17.707	**0.734**	0.225	0.891	0.642
Our-AttLWB	**17.783**	0.726	**0.220**	**0.896**	**0.706**

4.5 Results of Human Novel View Synthesis

Evaluation Metrics. As for data in the iPER dataset, we have videos containing different views of a certain subject performing A-pose, and in the MotionSynthetic dataset, we render A-pose images with 3D meshes from different viewpoints. Thus, we obtain images of the same person in different views. For evaluation, we use PSNR, SSIM [62] and LPIPS [63] as the metrics.

TABLE 4: Results for human novel view synthesis of different methods, including AppFlow [41], MV2NV [69], ours LWB and AttLWB, on iPER and MotionSynthetic dataset. Here, we report the PSNR, SSIM and LPIPS [63]. ↑ means the larger the better. A higher SSIM may not mean a better quality of an image [63].

	iPER			MotionSynthetic		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
AppFlow	23.342	0.849	0.133	25.575	0.896	0.083
MV2NV	24.950	**0.883**	0.125	25.951	0.837	0.097
LWB	24.518	0.862	0.090	25.055	0.779	0.106
AttLWB	**25.246**	0.867	**0.078**	**28.625**	**0.934**	**0.037**

Quantitative Results. In Table 4, we report the results of our methods AttLWB and that of other state-of-the-art methods, including AppFlow [41] and MV2NV [69], on the iPER and MotionSynthetic datasets based on the above evaluation metrics. The results show that our method outperforms other methods.

Qualitative Results. We randomly sample source images from the testing set of the iPER dataset and change the views from 30° to 330°. The results are illustrated in Fig. 12. Our method is capable of predicting reasonable content of invisible parts when switching to other views and keep the source information, in terms of face identity and clothes details, even in the self-occlusion case,

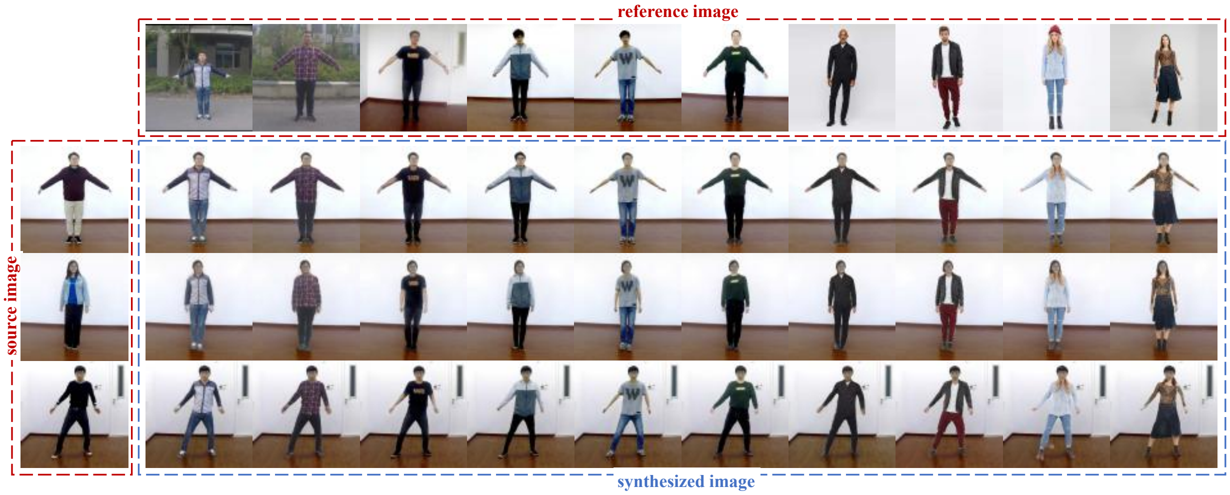


Fig. 11: Examples of our proposed AttLWB of human appearance transfer in the testing set of iPER (zoom-in for the best of view). All results are in 512×512 resolution. Our method could produce high-fidelity and decent images that preserve the face identity and shape consistency of the source image and keep the clothes details of reference image.

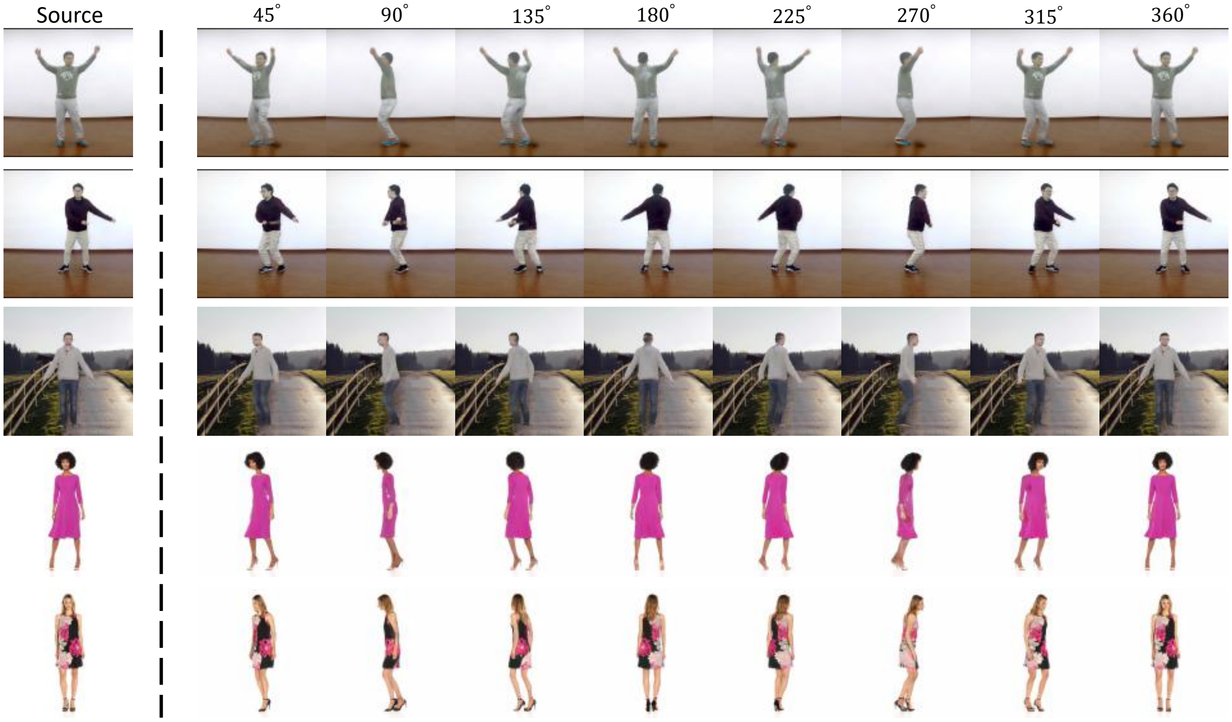


Fig. 12: Examples of our proposed AttLWB of human novel view synthesis. It is capable of preserving the source information, in terms of face identity and logo details of cloths, even the person wearing the long dress with fluffy hair.

such as the middle and bottom rows in Fig. 12. Through Fig. 12, we can see that 1) even when the subjects have large motion deformation, such as the case in the 1^{st} row of Fig. 12, results of our method can keep the logo details of clothes. 2) The 2^{nd} row shows the results when the subjects have self-occlusion. 3) Our method can also handle cases with complex background as the 3^{rd} row in Fig. 12 shows. 4) The 4^{th} row of Fig. 12 shows cases in which subjects wear a long dress and have fluffy hair. 5) The 5^{th} row of Fig. 12 is the case with complex clothes texture.

5 Ablation Studies and Analysis

In this section, we perform experiments to analyze the impacts of factors in our system, including with/without personalization, ablation studies of different loss functions and the comparison of our proposed LWB or AttLWB with other warping strategies, such as input concatenation, texture warping and feature warping. We further report the running time and analyze the failure cases.

5.1 Impact of Personalization

We perform the ablation studies of with/without personalization to verify the effectiveness of personalization. Besides, we also

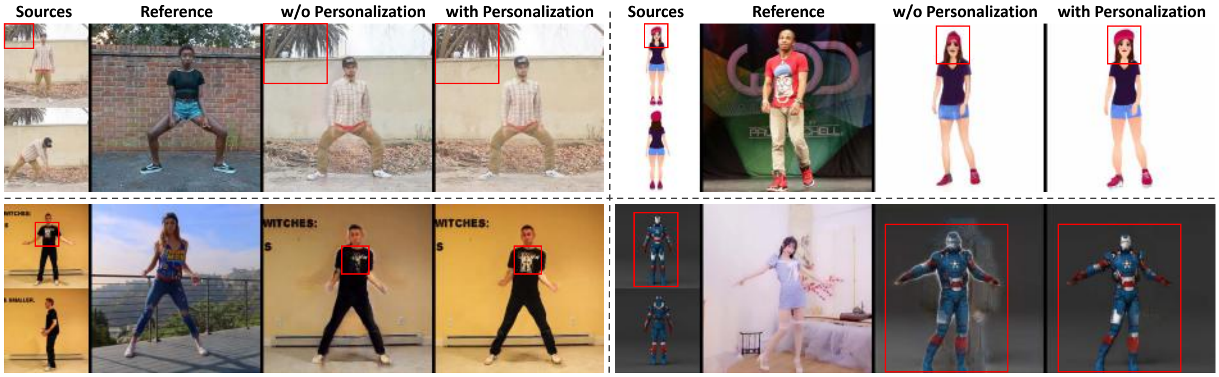


Fig. 13: Comparison of our proposed AttLWB with and without personalization (zoom-in for the best of view). The two roles in the left column are the source images from the Youtube-Dancer-18 dataset, and that in the right column are cartoon images from the Internet. From the top left segment, we can see that our method could preserve the color style of the source background with personalization. From the bottom-left segment, we find that our method without personalization might lose the details of the logo structure in the source images, while our method with personalization could preserve the logo details. The roles in the right column demonstrate that with personalization, our method has more capability of generalization; the results show that our model can deal with scenarios in which the source images are out of the domain of training set and even when the source images are in cartoon style from the Internet.

analyze the effect of hyper-parameters, including the number of source images $1 \leq s_n \leq 8$ and that of steps T for personalization. Since we only use the Youtube-Dancer-18 dataset in the testing phase, it is reasonable to evaluate the generalization of our methods of with/without personalization on this dataset. Here, we use self-imitation evaluation metrics, as mentioned above.

TABLE 5: Comparison of our proposed AttLWB with and without personalization on the Youtube-Dancer-18 dataset. ↑ means the larger the better and ↓ means the smaller the better.

	PSRN↑	SSIM↑	LPIPS↓	Body-CS↑	Face-CS↑	FID↓
w/o	16.932	0.519	0.302	0.792	0.335	79.321
with	**17.974**	**0.579**	**0.263**	**0.834**	**0.413**	**59.832**

With/Without Personalization. We conduct comparative experiments with and without personalization in our methods. Here, we fix the $s_n = 2$ and $T = 100$ in the phase of personalization. Table 5 shows that our method with personalization could achieve 1.0421 higher in PSNR, 0.0599 higher in SSIM, and 0.039 lower in LPIPS than that without personalization on the Youtube-Dancer-18 dataset. Furthermore, we display some example results in Fig. 13, where the left-column two roles are the source images from the Youtube-Dancer-18 dataset, and the right-column two roles are the cartoon images from the Internet. We find that with personalization, 1) our method could keep the color style of the background unchanged, as shown in the 1^{st} top left of Fig. 13; 2) our method is capable of preserving the logo details in the source clothes, as depicted in the 2^{nd} bottom left of Fig. 13; 3) our method is more powerful in the generalization, even when the source images are cartoon style, as illustrated in the right column of Fig. 13. These demonstrate that personalization indeed plays a significant role in improving the generalization of our system.

Number of Source Images s_n. In our system, we adopt a few source images $1 \leq s_n \leq 8$ for personalization, and we will analyze the impacts of s_n to the final results. Here, we fix the number of steps to $T = t \in \{10, 50, 100, 150, 200\}$ respectively for personalization and list the PSNR with different $s_n \in \{1, 2, 4, 8\}$ in Fig. 14. It shows that the performance grows with an increase of s_n when T is large enough. The reason for the performance increase is due to the increase of the invisible textures. However, it is worth noticing that when T is small with respect to a large s_n, in the case of $T = 10$ and $s_n = 8$, the performance decreases. The reason might be that when T is small, it is too hard for the network to fit those too many source images.

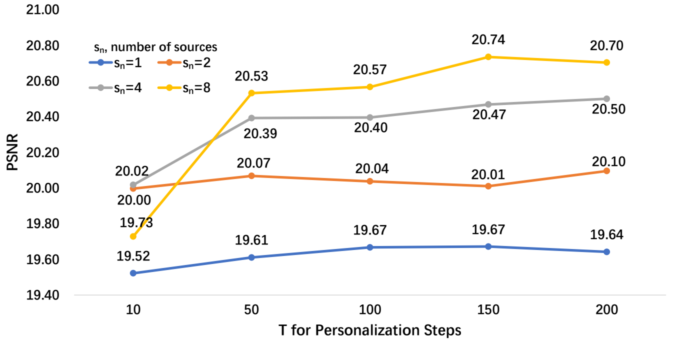


Fig. 14: Comparison of different number of source images s_n and number of steps T for personalization. The performance grows with the increase of s_n, when T is large enough. When T is small with respective to a large s_n, in this case of $T = 10$ and $s_n = 8$, the performance would decrease.

Number of Steps T for Personalization. In the real application, we should take the number of steps T into consideration because more steps will take more time. It is necessary to consider the trade-off between performance and overhead time for personalization. We set $s_n \in \{1, 2, 4, 8\}$ and list the performance with different T for personalization in Fig. 14. From Fig. 14, we can see that the performance saturates at around 150 steps.

In summary, based on the above analysis, we recommend that in the stage of personalization, finetuning around 100 steps should be enough, and if the time for personalization is limited, it would be better to use fewer source images.

5.2 Impact of Different Loss Functions

In our methods, we apply a perceptual loss L_p, a face identity loss L_f, an attention regularization loss L_a, and an adversarial loss L^G_{adv} (with global, body and head adversarial loss in details) to the full training loss functions. To validate the effectiveness of each term, we perform the ablation studies of the different loss functions. From Table 6, we can see that the model with the full loss would have the best performance. Besides, with the addition of L_f and L^G_{adv}, the performance increases compared with that of the trial with only L_p.

TABLE 6: Comparison between results with different loss functions on the Youtube-Dancer-18 dataset.↑ means the larger the better and ↓ means the smaller the better.

	PSNR↑	SSIM↑	LPIPS↓	Body-CS↑	Face-CS↑
L_p	18.204	0.575	0.274	0.791	0.314
$L_p + L^G_{adv}$	19.656	0.638	0.231	0.810	0.334
$L_p + L^G_{adv} + L_f$	19.542	0.629	0.247	0.809	0.351
L_{full}	**20.038**	**0.656**	**0.212**	**0.826**	**0.421**

5.3 Impact of Different Warping Strategies

To verify the impact of our proposed Attentional Liquid Warping Block (AttLWB), we design some baselines with the ways mentioned above to propagate the source information, including input concatenation, texture warping, and feature warping. The body recovery, flow composition modules, the basic network architectures, and all loss functions are the same except for the propagating strategies among our method and other warping baselines. Here, we denote early concatenation, texture warping, and feature warping, as W_C, W_T, and W_F, respectively. Also, we denote the s_n source images as $\{I_{s_1}, ..., I_{s_n}\}$, their corresponding conditional inputs as $\{C_{s_1}, ..., C_{s_n}\}$ and their corresponding feature maps as $\{X^l_{s_1}, ..., X^l_{s_n}\}$ at the l^{th} layer, respectively. The reference conditional inputs are C_t. The transformation flow of each source image to the reference is $T_{s_i \to t}$. We list the details of all warping baselines in followings:

Input Concatenation W_C. It directly concatenates all source images, their corresponding conditional inputs, as well as the reference conditional inputs, and then feeds them into the G_{TSF} network, as shown in Fig. 2 (a).

Texture Warping W_T. Based on each transformation flow $T_{s_i \to t}$, we warp each source image s_i to the reference condition, average the pixels of overlap regions, and synthesize an initial image. Then, we feed it into the G_{TSF} network and generate the final image, as shown in Fig. 2 (b).

Feature Warping W_F. Instead of warping the source information in the image space, it propagates the source information in the feature space, based on the transformation flow. As mentioned above, we firstly obtain the warped feature $X^l_{s_i \to t}$ by using a bilinear sampler (BS) to warp each source feature $X^l_{s_i}$ concerning the corresponding transformation flow $T_{s_i \to t}$. According to the ways to aggregate the global feature X^l_t from multiple warped source features $\{X^l_{s_1 \to t}, ..., X^l_{s_n \to t}\}$, we can specifically subdivide them into the followings:

1) **Attention** W^{Att}_F (ours) is shown in Algorithm 2.
2) **Add-Aggregation** W^A_F (ours). It is the first version of our proposed Liquid Warping Block(LWB) [21], as shown in the Fig. 4 (a) and Equation (2).
3) **Mean-Aggregation** W^M_F. Directly adding the warped features will enlarge the magnitude of the features in the overlap area and thereby results in artifacts. A naive way is to average all the warped features, shown as follows.

$$\hat{X}^l_t = \frac{1}{s_n} \sum_{i=1}^{s_n} X^l_{s_i \to t} + X^l_t. \tag{9}$$

4) **Add-Soft-Gate** $W^{A\odot}_F$. The warped feature might introduce the misalignment problem, and to address it, Dong *et al.* [11] utilizes a gated convolution to control the transformation degree. We firstly add all the warped features, then utilize a gated convolution, as shown in Equation (10). Here, g is a function with two-convolution layers followed by a Sigmoid activation and $g(X^l_t) \in [0, 1]$. $\odot$ represents the element-wise multiplication.

$$\hat{X}^l_t = g(X^l_t) \odot \sum_{i=1}^{s_n} X^l_{s_i \to t} + X^l_t. \tag{10}$$

5) **Mean-Soft-Gate** $W^{M\odot}_F$. It firstly averages all the warped features and following steps are the same with $W^{A\odot}_F$. The formulation is shown as follows:

$$\hat{X}^l_t = g(X^l_t) \odot \frac{1}{s_n} \sum_{i=1}^{s_n} X^l_{s_i \to t} + X^l_t. \tag{11}$$

We conduct a user study, with 64 volunteers, to assess the quality of the generated videos and compare the performance of the warping strategies mentioned above. Participants are shown 17 groups of videos with 7 videos generated by 7 warping strategies respectively in random order in each group. Besides, the shared source image and reference video of each group is also shown to the participants for reference. Participants are asked to choose the best video considering the quality of the face, clothes texture, figure pose, and background. Finally, 64 responses are collected, and the results are shown in Fig. 15. As we can see that our proposed AttLWB and AddLWB have the best appraise, scoring 41.73% and 20.04%, respectively, far higher than others.

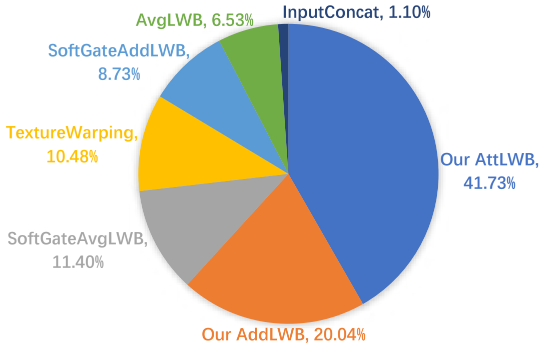


Fig. 15: Results of the user study (%). The user preference of the videos with best quality regarding to the quality of face, the quality of clothes texture and background.

5.4 Running Time

Our method could produce the results with different image resolutions, ranging from 256×256, 512×512, 1024×1024 to 1920×1920. Here, we benchmark the running time of our system in different image resolutions. Since a high resolution needs more memory allocation of the GPUs device, we perform all the tests on a Tesla V100S-PCIe-32G GPU with the Intel Xeon(R) E5-2620 2.10GHz CPUs. The image resolution of the source images is 4032×3024, and that of the reference video with 165 frames is 1920×1080. In Fig.16, we separately report the running time of preprocessing, personalization and inference, when synthesizing different resolutions, respectively. From Fig.16, we can see that the higher resolution consumes more running time, especially in the personalization and inference.

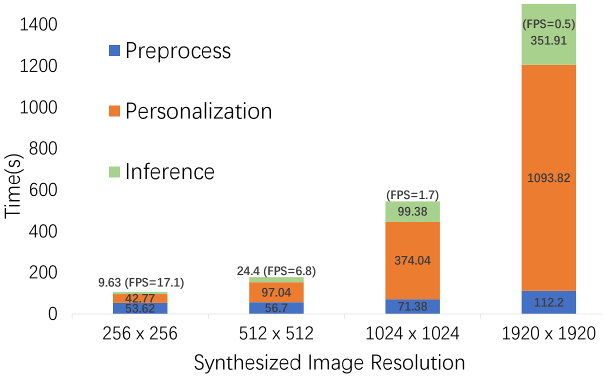


Fig. 16: Running time when producing images with different resolutions. The I/O consumption has been taken into count. The larger resolution, the more consuming time is, particularly in the stages of personalization and inference.

5.5 Failure Cases and Limitations

There are three main types of failure cases of our methods. The first one, as shown in the 1^{st} row of Fig. 17, is that source image contains a large area of self-occlusion, which introduces an ambiguity in textures and thereby results in a bad synthesized image. The second occurs when the Body Recovery Module fails and could not accurately estimate the pose parameters, as illustrated in the 2^{nd} row of Fig. 17. The rest is when the background inpaintor G_{BG} fails, as shown in the 3^{rd} row of Fig. 17.

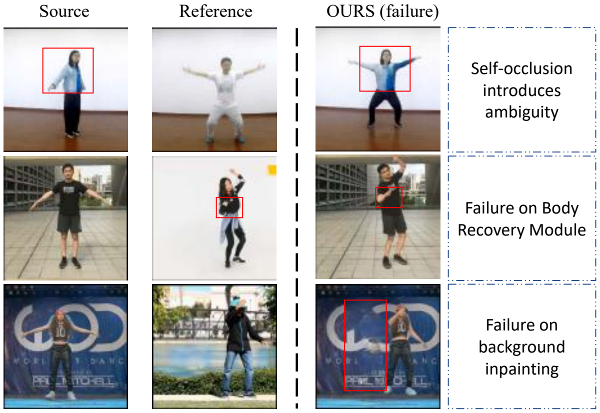

Source
Reference
OURS (failure)
Self-occlusion introduces ambiguity
Failure on Body Recovery Module
Failure on background inpainting

Fig. 17: The failure cases of our system. It mainly contains three types of failure cases. One occurs when the source images introduce a large self-occlusion area, as shown in the top row. The second row is when the body recovery module fails. The third row shows the artifacts when the background inpainting network fails.

In addition, there are still some limitations of our system, 1) it cannot imitate the motions of hands and facial expressions from the reference images, since the 3D body parametric SMPL [16] used in our system does not contain the articulated hands and expressive face; 2) also, it cannot animate the large-motion body with too loose clothing like the skirt or evening dress; 3) it is affected by the different lighting environments among sources.

Therefore, for a better result, the input source images need to follow these guidelines:

- They share the same static background without too complex scene structures. If possible, we recommend using the actual background.
- The person in the source images holds an A-pose for introducing the most visible textures.
- It is recommended to capture the source images in an environment without too much contrast in lighting conditions and lock auto-exposure and auto-focus of the camera.

6 Conclusion

We propose a unified framework to handle human motion imitation, appearance transfer, and novel view synthesis. It employs a body recovery module to estimate the 3D body mesh, which is more powerful than the 2D poses. In order to preserve the source information, we further design a novel warping strategy, Attentional Liquid Warping Block (AttLWB), which propagates the source information in both image and feature spaces and supports a more flexible warping from multiple sources. Besides, with a fast personalization, our method could be generalized well when the input images are out of the domain of training set and synthesize higher resolution (512×512 and 1024×1024) results. Extensive experiments show that our framework outperforms others and produce decent results.

Acknowledgment

We thank Dr. Weixin Luo for the meaningful discussion in the whole procedure, and we appreciate all the help of building the first version of the iPER dataset from Min Jie.

References

[1] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag, “Synthesizing images of humans in unseen poses,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[2] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, “Pose guided person image generation,” in *Advances in Neural Information Processing Systems*, 2017, pp. 405–415.

[3] A. Siarohin, E. Sangineto, S. Lathuilière, and N. Sebe, “Deformable gans for pose-based human image generation,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[4] A. Raj, P. Sangkloy, H. Chang, J. Hays, D. Ceylan, and J. Lu, “Swapnet: Image based garment transfer,” in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XII*, 2018, pp. 679–695.

[5] M. Zanfir, A.-I. Popa, A. Zanfir, and C. Sminchisescu, “Human appearance transfer,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[6] B. Zhao, X. Wu, Z. Cheng, H. Liu, Z. Jie, and J. Feng, “Multi-view image generation from a single-view,” in *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, 2018, pp. 383–391.

[7] H. Zhu, H. Su, P. Wang, X. Cao, and R. Yang, “View extrapolation of human body from a single image,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[8] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976, 2017.

[9] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, “Everybody dance now,” in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[10] J. Lee, D. Ramanan, and R. Girdhar, “Metapix: Few-shot video retargeting,” in *ICLR*, 2019.

[11] H. Dong, X. Liang, K. Gong, H. Lai, J. Zhu, and J. Yin, “Soft-gated warping-gan for pose-guided person image synthesis,” in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, 2018, pp. 472–482.

[12] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, “Spatial transformer networks,” in *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, 2015, pp. 2017–2025.

[13] B. AlBahar and J.-B. Huang, “Guided image-to-image translation with bi-directional feature transformation,” in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[14] N. Neverova, R. A. Güler, and I. Kokkinos, “Dense pose transfer,” in *European Conference on Computer Vision (ECCV)*, 2018.

[15] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, “Keep it smpl: Automatic estimation of 3d human pose and shape from a single image,” in *European Conference on Computer Vision*. Springer, 2016, pp. 561–578.

[16] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, “SMPL: A skinned multi-person linear model,” *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, vol. 34, no. 6, pp. 248:1–248:16, oct 2015.

[17] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, “End-to-end recovery of human shape and pose,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[18] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, “Learning to reconstruct 3d human pose and shape via model-fitting in the loop,” in *ICCV*, 2019.

[19] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, “Deepfashion: Powering robust clothes recognition and retrieval with rich annotations,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[20] P. Zablotskaia, A. Siarohin, B. Zhao, and L. Sigal, “Dwnet: Dense warp-based network for pose-guided human video generation,” in *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*. BMVA Press, 2019, p. 51.

[21] W. Liu, Z. Piao, J. Min, W. Luo, L. Ma, and S. Gao, “Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis,” in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, 2017, pp. 5998–6008.

[23] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, “Semantic image synthesis with spatially-adaptive normalization,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[24] T. R. Shaham, T. Dekel, and T. Michaeli, “Singan: Learning a generative model from a single natural image,” in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[25] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, “Few-shot adversarial learning of realistic neural talking head models,” in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[26] P. Esser, E. Sutter, and B. Ommer, “A variational u-net for conditional appearance and shape generation,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[27] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, “High-resolution image synthesis and semantic manipulation with conditional gans,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[28] T. Wang, M. Liu, J. Zhu, N. Yakovenko, A. Tao, J. Kautz, and B. Catanzaro, “Video-to-video synthesis,” in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, 20148, pp. 1152–1164.

[29] A. Shysheya, E. Zakharov, K.-A. Aliev, R. Bashirov, E. Burkov, K. Iskakov, A. Ivakhnenko, Y. Malkov, I. Pasechnik, D. Ulyanov, A. Vakhitov, and V. Lempitsky, “Textured neural avatars,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[30] L. Liu, W. Xu, M. Zollhoefer, H. Kim, F. Bernard, M. Habermann, W. Wang, and C. Theobalt, “Neural rendering and reenactment of human actor videos,” *ACM Transactions on Graphics 2019 (TOG)*, 2019.

[31] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz, “Disentangled person image generation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[32] C. Si, W. Wang, L. Wang, and T. Tan, “Multistage adversarial losses for pose-based human image synthesis,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[33] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, 2015, pp. 234–241.

[34] I. K. Rıza Alp Güler, Natalia Neverova, “Densepose: Dense human pose estimation in the wild,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[35] Z. Zhu, T. Huang, B. Shi, M. Yu, B. Wang, and X. Bai, “Progressive pose attention transfer for person image generation,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[36] Y. Li, C. Huang, and C. C. Loy, “Dense intrinsic appearance flow for human pose transfer,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[37] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black, “Clothcap: seamless 4d clothing capture and retargeting,” *ACM Trans. Graph.*, vol. 36, no. 4, pp. 73:1–73:15, 2017.

[38] C. Zhang, S. Pujades, M. J. Black, and G. Pons-Moll, “Detailed, accurate, human shape estimation from clothed 3d scan sequences,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 5484–5493.

[39] V. Leroy, J. Franco, and E. Boyer, “Multi-view dynamic shape refinement using local temporal integration,” in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2003, pp. 3113–3122.

[40] B. L. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll, “Multi-garment net: Learning to dress 3d people from images,” in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[41] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros, “View synthesis by appearance flow,” in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, 2016, pp. 286–301.

[42] E. Park, J. Yang, E. Yumer, D. Ceylan, and A. C. Berg, “Transformation-grounded image generation network for novel 3d view synthesis,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[43] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2672–2680.

[44] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, “Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization,” in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[45] Z. Ding, Y. Guo, L. Zhang, and Y. Fu, “One-shot face recognition via generative learning,” in *13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi’an, China, May 15-19, 2018*. IEEE Computer Society, 2018, pp. 1–7.

[46] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 2017, pp. 1126–1135.

[47] A. Nichol, J. Achiam, and J. Schulman, “On first-order meta-learning algorithms,” *CoRR*, vol. abs/1803.02999, 2018.

[48] D. Lian, Y. Zheng, Y. Xu, Y. Lu, L. Lin, P. Zhao, J. Huang, and S. Gao, “Towards fast adaptation of neural architectures with meta learning,” in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[49] T. Wang, M. Liu, A. Tao, G. Liu, B. Catanzaro, and J. Kautz, “Few-shot video-to-video synthesis,” in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, 2019, pp. 5014–5025.

[50] K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” in *Computer Vision - ECCV 2016 - 14th European Confer-*

ence, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV, 2016, pp. 630–645.

[51] H. Kato, Y. Ushiku, and T. Harada, "Neural 3d mesh renderer," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 2018, pp. 3907–3916.

[52] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016, pp. 770–778.

[54] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, 2016, pp. 694–711.

[55] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA*, May 2015.

[56] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[57] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 6738–6746.

[58] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "On the effectiveness of least squares generative adversarial networks," *CoRR*, vol. abs/1712.06391, 2017. [Online]. Available: http://arxiv.org/abs/1712.06391

[59] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "Ganimation: Anatomically-aware facial animation from a single image," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X*, 2018, pp. 835–851.

[60] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, "Video based reconstruction of 3d people models," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[61] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, vol. abs/1412.6980, 2015.

[62] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[63] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[64] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 779–788.

[65] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[66] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, 2016.

[67] https://github.com/timesler/facenet-pytorch/.

[68] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems 30: 4-9 December 2017, Long Beach, CA, USA*, 2017, pp. 6626–6637.

[69] S.-H. Sun, M. Huh, Y.-H. Liao, N. Zhang, and J. J. Lim, "Multi-view to novel view: Synthesizing novel views with self-learned confidence," in *European Conference on Computer Vision*, 2018.

Wen Liu received the bachelor degree from Northwestern Polytechnical University, Xian, China, in 2016. He is currently pursuing a Ph.D. degree at ShanghaiTech University. His research interests focus on human 3D body reconstruction, image synthesis, motion transfer, novel view synthesis, neural rendering and video anomaly detection.

Zhixin Piao received the bachelor degree from Southeast University, Nanjing, China, in 2017. He is currently pursuing a master degree at ShanghaiTech University. His research topic is human 3D reconstruction and motion transfer.

Zhi Tu received the bachelor degree from ShanghaiTech University, Shanghai, China, in 2020. His research topic is human motion transfer and medical image analysis.

Wenhan Luo received the Ph.D. degree from Imperial College London, UK, 2016, M.E. degree from Institute of Automation, Chinese Academy of Sciences, China, 2012 and B.E. degree from Huazhong University of Science and Technology, China, 2009. His research interests include several topics in computer vision and machine learning, such as motion analysis (especially object tracking), image/video quality restoration, object detection and recognition, reinforcement learning.

Lin Ma received the B.E. and M.E. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2006 and 2008, respectively, and the Ph.D. degree from the Department of Electronic Engineering, The Chinese University of Hong Kong, in 2013. He was a Researcher with the Huawei Noah's Ark Laboratory, Hong Kong, from 2013 to 2016. He is currently a Principal Researcher with the Tencent AI Laboratory, Shenzhen, China. His current research interests lie in the areas of computer vision, multimodal deep learning, specifically for image and language, image/video understanding, and quality assessment. Dr. Ma received the Best Paper Award from the Pacific-Rim Conference on Multimedia in 2008. He was a recipient of the Microsoft Research Asia Fellowship in 2011. He was a finalist in HKIS Young Scientist Award in engineering science in 2012.

Shenghua Gao is an assistant professor, PI in ShanghaiTech University, China. He received the B.E. degree from the University of Science and Technology of China in 2008 (outstanding graduates), and received the Ph.D. degree from the Nanyang Technological University in 2012. From Jun 2012 to Jul 2014, he worked as a postdoctoral fellow in Advanced Digital Sciences Center, Singapore. His research interests include computer vision and machine learning.