Single-Shot Freestyle Dance Reenactment

Oran Gafni
Facebook AI Research
oran@fb.com

Oron Ashual
Facebook AI Research
oron@fb.com

Lior Wolf
Facebook AI Research
and Tel-Aviv University
wolf@fb.com

(a) (b) (c)

Figure 1. Single-shot dance reenactment. Using only a single image of a target person and their corresponding extracted semantic map (a), and a driving person's pose (b), we are able to render a novel corresponding semantic map of the target person, and a realistic person in the novel pose (c). Unlike previous work, we are able to maintain the body shape of the target person.

Abstract

The task of motion transfer between a source dancer and a target person is a special case of the pose transfer problem, in which the target person changes their pose in accordance with the motions of the dancer. In this work, we propose a novel method that can reanimate a single image by arbitrary video sequences, unseen during training. The method combines three networks: (i) a segmentation-mapping network, (ii) a realistic frame-rendering network, and (iii) a face refinement network. By separating this task into three stages, we are able to attain a novel sequence of realistic frames, capturing natural motion and appearance. Our method obtains significantly better visual quality than previous methods and is able to animate diverse body types and appearances, which are captured in challenging poses.

1. Introduction

The goal of this work is to animate a target person, who is specified by a single input image, to mimic the motion of a driving person, who is captured in a video sequence. This pair of inputs can be considered the easiest to obtain, and most minimalist and generic input for the given synthesis problem. Importantly: both the input image and the driving video are unseen during training.

The method we propose extends the envelope of the current possibilities in multiple ways: (i) the target person can vary in body shape, age, ethnicity, gender, pose, and viewpoint (ii) the sequence of poses that form the motion can be unconstrained, which is why we emphasize freestyle dance, (iii) the background can vary arbitrarily and is not limited to the source image or the background of the driving video.

This general setting contrasts with the limitations of existing methods, which often struggle to maintain the target person's appearance and avoid mixing elements from the driving video. The existing methods also often require an input video of the target person, have difficulty producing natural motion, and are limited to specific backgrounds. This is true, even for methods that train to map between specific persons seen during training.

To achieve this novel set of capabilities, we make extensive use of the latest achievements of neural networks for human capturing. Two pre-trained pose recognition networks are used to analyze the input video, a pre-trained human parsing network is used to segment the input image (of the target person), a pre-trained face embedding network is

used to improve the face, and an inpainting network is utilized to extract the background of each training image. This maximal use of existing tools is an enabler for our method: using just one of the pose networks, or using pose in lieu of human parsing fails to deliver the desired results.

In addition to these components, for which there exist previous works that include a subset of it, we further employ specific representations. In order to ensure that the clothing and face appearance are captured realistically, we employ a five-part human encoder to the realistic frame-rendering network, consisting of four ImageNet-trained classifiers, and a trained face embedding network. These provide a rich embedding of the target, later enforced by a set of relevant perceptual losses. To ensure that finger motion is natural and the rendered hands do not suffer from missing parts, hand training data is augmented.

The method separates the pose and frame generation parts, performing each by a different network. The pose is provided in the space of a part-based segmentation map and is conditioned on both the target person and the motion frame. The second network transforms the generated pose and the target person's details to a masked frame, which is blended with an arbitrary background. The frame is further improved by applying a face refinement network based on an appearance preserving perceptual loss.

An extensive set of experiments is provided to establish the visual and numerical validity of the method. Compared to previous methods, our method provides considerably more accurate and visually pleasing results, as evaluated by a set of numerical metrics, a user study, and visual examples. Contrary to most previous work, we emphasize the ability to handle diversity in the target and generated individuals, promoting inclusion, which is generally lacking in this line of work.

2. Related work

A similar setting was presented in few-shot vid2vid (fsV2V) [41], which generates a video sequence given a driver video and a source image containing a target person. Like our method, this method only trained once and can then be applied to any pairs of inputs. However, there are major differences in the applicability of the methods: our method can generate in arbitrary backgrounds, broader ranges of motions and is less restricting with respect to the inputs. Technically, fsV2V employs a hypernetwork [18] that predicts the weights of a vid2vid network [42] given the target domain image(s), while our method employs conditioning based on this input. fsV2V suffers from flow-based artifacts, since it warps between consecutive frames, while our method generates entirely de-novo images.

DwNet [36] also warps the input image based on the motion of the driver video. Therefore, it is bound to the static background of the target person and suffers from artifacts around the animated character.

"Everybody dance now" [5] and vid2vid [42], similarly to [41] generate an entire image, which includes both the target character and its background, resulting in artifacts near the edges of the generated pose [33, 6], background motion artifacts, and blurriness in some parts of the background. We employ a mask-based solution to integrate the generated character into an arbitrary background. Masks were previously used in the context of dancing to reanimate a specific person [56]. Methods that model a specific person do not need to model variation in body shape or capture novel appearances from a single frame.

Unlike our work and fsV2V, many methods require the target person to be specified by a video containing sufficiently varied motion (and not just an arbitrary still image), and are retrained per each pair of motion-source video and target-person video [5, 43, 47, 35].

vid2game [16] is also trained per-person on a video containing a character's motion. Another difference from our work is that there is no replacement of appearance nor transfer of motion. Similar to our work, vid2game incorporates two networks Pose2Pose (P2P) and Pose2Frame (P2F), which are analog to two of the networks we use. However, the inputs and outputs differ from those of our networks, and the P2P network of vid2game generates similar poses in an autoregressive manner, while our task is more related to pose transfer. While vid2game is trained in a fully supervised manner, our network is trained in a self-supervised manner to reconstruct a person that exists in the image.

Once the frame is obtained, we employ a face refinement network that utilizes an autoencoder architecture similar to the de-ID network [14]. While [14] seeks to distance the appearance from that of a target person, our method has opposite goals, bringing the appearance closer.

In still images, the problem of pose transfer is well studied [28, 37, 1, 45, 9, 57, 11, 10, 39, 30, 26], out of which [9, 39, 30] use a human parser, as we do. Most of these contributions employ images from the DeepFashion dataset [27], which has four prominent disadvantages. First, the images are set against a white background; second, the poses are limited to those encountered in fashion photography, and for example, the hands are rarely above the head; third, the body shapes in the dataset are limited, and fourth, the number of different appearances, ethnicities and ages are few, resulting in overfitting to specific gender and age types. Another popular benchmark is the Market-1501 dataset [53], which depicts low-resolution images, with limited pose variability, that greatly differ from the dancing reenactment scenario. Explicit 3D modeling for single-image reanimation has been practiced as well [46], yet tends to result in unnatural motion and suffers from artifacts resulting from target image occlusions.

3. Method

Our method reenacts a character specified by a single input image, based on a given sequence of pose-frames. The method is designed to be generic, and the models are trained once and can then be applied, at test time, to any input character and motion sequence, without adjustments, re-training, or fine-tuning.

The method relies on three image2image networks, each trained independently: (i) the P2B (Pose-to-Body) network maps pose and character information into body data, (ii) the B2F (Body-to-Frame) network maps the body-pose information obtained from the B2P and the character information to a frame, and (iii) the FR network refines the face in the frame generated by the P2F network.

On top of the three main networks we train (P2B, B2F, and FR), we employ an extensive set of pre-trained networks, in a manner that is unprecedented as far as we can ascertain: (i) a VGG network [38] trained on the ImageNet [7] dataset that is used for obtaining the perceptual loss while training the B2F. (ii) A face detection and 2D alignment network [2]. (iii) VGGFace2, which is a face embedding network [3] that is used for training both the B2F and FR networks. (iv) The DensePose [34] network and (v) the OpenPose [4] network are both used to obtain pose information from each frame, as a way to represent the input of P2B. (vi) A human parsing network HP [24] is used to extract the body in the target image. (vii) An inpainting network [49, 48] extracts the background from the training images, as well as from the target image at inference time.

During training, we employ additional networks as discriminators that are denoted by D_k. There are a total of five discriminators: two are used for training the P2B, two for training the B2F, and one for training the FR.

The index $i = 1, 2, ..$ is used to denote a frame index. The generated video frames (constructed from the output of B2F and FR) are denoted by f_i. The output of P2B is a sequence of generated semantic maps P_i^M that are trained to mimic the output HP provides on real images of human figures. The input to P2B is comprised of two sequences: P_i^D and P_i^S, denoting the dense annotation provided by DensePose, given a video v and the stick figure and face landmarks output of OpenPose on v, respectively. In addition, P2B receives a semantic map p^{M*} that denotes the parsing obtained by network HP for an input image I, that is used to specify the (target) person to reenact.

B2F receives as input the sequence P^M (here and below, the index is omitted to denote the entire sequence) and e_z, which is the concatenated embedding extracted by the pre-trained VGGFace2 and VGG encapsulating the target person appearance. The output of P2F consists of two sequences: z_i denotes the generated image information, and m_i is a sequence of blending masks (values between 0 and 1), that determines which image regions in the frame output would contain the information in z_i and which would contain the background information provided by the user. The background information is denoted by b_i and can be dynamic. The combination of the background with the synthesized images, in accordance with the masks is denoted by f_i^0. The output frames are generated by applying the refinement network FR to it.

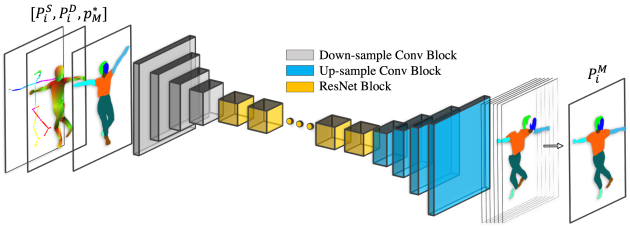


Figure 2. The architecture of the P2B network. Given a semantic segmentation of the target body p^{M*}, a source pose P_i^S, and a source dense pose P_i^D, the network generates the semantic map P_i^M of the target person in the desired pose.

Our method's flow consists of the following set of equations, given the input sequence of background frames b, image specification of the target person I, and a video containing the desired motion v.

$$p^{M*} = \mathrm{HP}(I) \tag{1}$$
$$P_i^D, P_i^S = \mathrm{DP}(v_i), \mathrm{OP}(v_i) \tag{2}$$
$$P_i^M = \mathrm{P2B}(p^{M*}, P_i^S, P_i^D) \tag{3}$$
$$t_1, t_{2-5} = \mathrm{l}(I, p^{M*}) \tag{4}$$
$$e_z = [\mathrm{VGGFace2}(t_1), \mathrm{VGG}(t_{2-5})] \tag{5}$$
$$(z_i, m_i) = \mathrm{B2F}(P_i^M, e_z)) \tag{6}$$
$$f_i^0 = z_i \cdot m_i + b_i \cdot (1 - m_i) \tag{7}$$
$$f_i = \mathrm{FR}(f_i^0, t_1) \tag{8}$$

where $i = 1, 2, ..$, HP, DP, and OP are the Human-Parsing, DensePose and OpenPose networks respectively, the $P2B$ and $B2F$ are the Pose2Body and the Body2Frame networks. l (Eq 6) is a function that separates the input image I into 5 stacked 224x224 images t_{1-5}, containing the appearance of the (1) face and hair, (2) upper-body clothing, (3) lower-body clothing, (4) shoes and socks, and (5) skin tone, in accordance with the semantic parsing map p^{M*}. As stated, $B2F$ returns a pair of outputs, an image z_i and a mask m_i that are linearly blended with the desired background b_i to create the initial frame f_i^0, using a per-element multiplication operator denoted by $(\cdot)$. FR takes this initial frame, and updates the face to better resemble the face of the target person, as captured in I. The semantic segmentation maps P_i^M and p_i^M are used in order to specify the face areas in the generated frame f_i^0 and in I, respectively.

3.1. Pose2Body network

The P2B's objective is to capture and transfer motion into the desired body structure, one frame at a time. The network has three inputs p^{M*}, P_i^S, and P_i^D. the first is produced by the human parser network applied to image I, the other two are obtained by pose networks, as applied to frame i of the motion-driving video. The parsing map p^{M*} consists of 22 labels, of which 20 labels are used as in the VIP dataset [54], and 2 labels are added to augment the hand landmarks extracted by OpenPose as labels.

DensePose outputs three channels of the UV(I) space, where two channels project 3D mapping to 2D, and the third is a body index channel, with values between $0-24$. OpenPose generates key-points, which are joined to a single RGB stick-figure. Facial and hand landmarks are added to the stick-figure, increasing certainty and stability to the generated output.

The P2B network utilizes the architecture of pix2pixHD [43]. In contrast to its original use for unconditioned image-to-image cross-domain mapping, we modify the architecture to allow it to generate a semantic segmentation map. Specifically, P2B produces the output P_i^M, which lies in the same domain as p^{M*}.

The architecture of P2B is illustrated in Fig. 2. Three inputs of the same spatial dimension are concatenated to one input tensor. The encoder part of the network is a CNN with ReLU [31] activations and batch normalization [21]. The latent space embedding goes through n_r residual blocks. Finally, the decoder u employs fractional strided convolutions [12], ReLU activations, and instance normalization [40]. A $\mathrm{sigmoid}$ non-linearity is applied after the last convolution to generate the output segmentation map.

3.1.1 Training the Pose2Body network

Following [43], we employ two discriminators (low-res and high-res), indexed by $k=1,2$. During training, the LSGAN [29] loss is applied. An L1 feature-matching loss is applied over both discriminators' activations. In contrast to the B2F implementation, we apply a cross-entropy loss over the generated output.

The loss applied to the generator can be formulated as:

$$\mathcal{L}_{P2B} = \sum_{k=1}^{2} \left(\mathcal{L}_{LS^k} + \lambda_D \mathcal{L}_{FM_D^k} \right) + \lambda_{CE} \mathcal{L}_{CE} \tag{9}$$

where the networks are trained with $\lambda_D = 40, \lambda_{CE} = 1$. The LSGAN generator loss is:

$$\mathcal{L}_{LS^k} = \mathbb{E}_{(p_i^A)} \left[\left(D_k(P2B(p_i^A)) - \mathbb{1} \right)^2 \right] \tag{10}$$

The expectation is computed per mini-batch, over the input HP, OP and DP $p_i^A = p^{M*}, P_i^S, P_i^D$. The discriminator-feature matching-loss compares the ground-truth semantic map with the generated one, using the activations of the discriminator, and is calculated as:

$$\mathcal{L}_{FM_D^k} = \mathbb{E}_{(p_i^A)} \sum_{j=1}^{M} \frac{1}{N_j} ||D_k^{(j)}(P_i^M) - D_k^{(j)}(P2B(p_i^A))||_1 \tag{11}$$

with M being the number of layers, N_j the number of elements in each layer, and $D_k^{(j)}$ the activations of discriminator k in layer j. The CE loss forces the generated 22 channels P_i^M to be similar to the ground truth semantic map P_i^{M*}, and can be formulated as:

$$\mathcal{L}_{CE} = CE(P_i^{M*}, P2B(p_i^A)) \tag{12}$$

P2B is trained using the Video instance-level Parsing (VIP) dataset [55]. The dataset provides semantic segmentation annotations of people in diverse scenarios. Each training step relies on a single person in two different poses. To segment individuals in different views and poses, we rely on their location in a random frame, and an additional random frame, limited to a range of 250 consecutive frames. From the first, we utilize the semantic annotation, and DP/OP (Eq (2)) as the network input, and the second is used for the semantic segmentation annotation ground truth, guiding towards the desired body-type and clothing.

Disentangling body structure. Few-shot generation methods suffer from the inability to generate a diverse set of body structures, as it is both challenging to correctly capture a body structure by a few samples, and datasets are highly biased towards certain body types. As a result, networks tend to learn a transformation of the source body structure, through the stick or dense pose representation, to the generated body structure.

In addition to data augmentation in the form of random rotation and scaling of the inputs and output, we establish a more robust form of disentanglement between the guiding poses P_i^S, P_i^D and the generated and source semantic maps p^{M*}, P_i^M , by introducing an additional form of data augmentation which is independent of the input and output body structures. We deliberately create a mismatch between the poses and semantic maps, by squeezing and stretching solely the body structures (segmentation maps) rather than the input poses. The network experiences samples that are in the exact same pose and view, yet differ in body structure. Examples of diverse body structure capability can be seen in Fig. 1 and in the supplementary.

3.2. Body2Frame network

B2F relies on two sources of input information: the generated pose of the target person P_i^M and the encoding of the target person's image I. The latter is obtained based on image I and its segmentation map p^{M*}. A stack

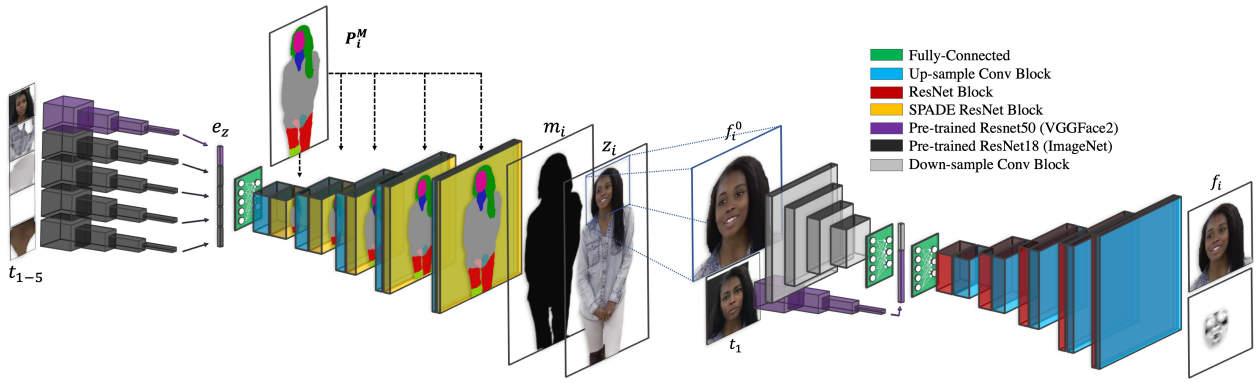


Figure 3. B2F and FR architectures. B2F receives as input the tensor $l(I, p^{M*})$ in which the segmented parts of the image I are introduced through an array of pre-trained networks, and a conditioning semantic map p_i^M. The output frame f_i^0 is generated by blending a generated frame z_i with the background b_i in accordance with a generated mask m_i. FR extracts a face embedding utilizing a trained face embedding network and concatenated to the latent space. The pose, expression and lighting conditions are encoded for each input frame by the encoder, while the appearance can be taken from any image of that person.

$t_{1-5} = l(I, p^{M*})$ of five 224x224 images is created, corresponding to the resized bounding boxes around five semantic segments: (1) face and hair, (2) upper-body clothing, (3) lower-body clothing, (4) shoes and socks, and (5) skin tone.

The output of B2F is a high-resolution (512×320) frame f_i^0. The frames in the sequence $i = 1, 2, ..$ are generated one by one, similarly to the P2B network. Each frame is generated by blending the background frame b_i (can be static or dynamic) with the two outputs of B2F, the mask m_i and the generated image z_i, as formulated in Eq. 7.

Architecture. The architecture of the B2F network is depicted in Fig. 3. Image t_1 is passed through a pre-trained face embedding network to extract the appearance embedding, while images t_{2-5} are encoded using a network pre-trained over the ImageNet dataset. The embedding extracted from the five pre-trained networks is concatenated into a single vector e_z of size $2048 + 4 * 512 = 4096$. The latent space is projected by a fully connected layer to obtain a vector that is a reshaped tensor of size $4 \times 4 \times 1024$. The decoder has seven upsample convolutional layers with interleaving SPADE [32] blocks. At test time, the latent space and FC layer are constant for a specific user, hence run only once, increasing the method's speed and applicability.

Datasets. To enable diverse generation capabilities in terms of appearance (ethnicity, gender and age), pose, and perspective, we combine the Multi-Human Parsing (MHPv2) [52, 22] and the Crowd Instance-level Human Parsing (CIHP) [17] datasets. Both datasets contain various poses, viewpoints, and appearances, increasing the robustness of the network. Every annotated person is cropped to provide a single sample, that is later randomly resized for data augmentation purposes.

Face emphasis. Although a face refinement network is applied to the B2F output, it is limited in its refinement capabilities. Therefore, the B2F is required to generate a high-quality face as part of the novel person. The desired target face is introduced through the embedding, as extracted by the trained face embedding network. To encourage the generated face to be similar to the target face, both in quality and appearance, we apply a set of perceptual losses aimed at the expected position of the generated face. This is done in a pre-processing step, where all face locations are calculated using the face annotation. During training, these locations are adjusted to the random transformations applied, such as resizing, cropping, and flipping.

We apply a perceptual loss over the low, mid and high-level activations of a trained face embedding network. While high-level abstractions encourage appearance preservation, lower-levels handle other aspects, such as expressions.

Additional guidance is provided to the face area in the form of explicit labels. Facial landmarks are used to draw five additional labels for the (1) eyebrows, (2) eyes, (3) nose, (4) lips, and (5) inner mouth. Although these landmarks are extracted from the driving (source) video, the perceptual losses applied to the face, as described in Eq. 18, help preserve the target person's appearance and expression.

Blending mask. B2F generates a blending mask in tandem with the generated character. This is imperative, as it enables the generated person to be embedded in any static or dynamic scene naturally. Training the B2F on an image dataset introduces an additional strain on the learning process of the blending mask, as there is no background image where the character is not present. To tackle this, we add a pre-processing step of inpainting all images, regenerating a region obtained by dilating the union of all semantic segmentation masks obtained by HP. To increase generation quality, all losses are applied solely to the character. The

semantic segmentation annotation labels are used to mask irrelevant image areas, such as the background or other people present in the same crop.

The following losses are used for training B2F:

$$\mathcal{L}^{G}_{hinge} = -\|D_{1,2}(P^M_i, z^b)\|_1 \tag{13}$$

$$\mathcal{L}^{D_{1,2}}_{hinge} = -\|\min(D_{1,2}(P^M_i, z^b) - 1, 0)\|_1 - \|\min(-D_{1,2}(P^M_i, x^b) - 1, 0)\|_1 \tag{14}$$

$$\mathcal{L}^{D_{k=1,2}}_{FM} = \mathbb{E}_{(P^M_i, x^b, z^b)} \sum_{j=1}^{M} \frac{1}{N_j} ||D^{(j)}_k(P^M_i, x^b) - D^{(j)}_k(P^M_i, z^b))||_1 \tag{15}$$

with M being the number of layers, N_j the number of elements in each layer, $D^{(j)}_k$ the activations of discriminator k in layer j, $z^b, x^b = z \odot P^{D+}_i, x \odot P^{D+}_i$, and $L^{G/D}_{hinge}$ as in [50, 25].

$$\mathcal{L}^{VGG}_{FM} = \sum_{j=1}^{M} \frac{1}{N'_j} ||VGG^{(j)}(x) - VGG^{(j)}(o))||_1 \tag{16}$$

with N'_j being the number of elements in the j-th layer, and $VGG^{(j)}$ the VGG classifier activations at the j-th layer.

The network also outputs a mask, which is trained using the L1 loss to reconstruct a binary version of the HP frame P^M_i after threshold at zero, denoted by P^{D+}_i ($\lambda_m = 5.0$):

$$\mathcal{L}^m_i = \lambda_m \|m_i - P^{D+}_i\|_1 \tag{17}$$

3.3. Face refinement network

The third network, FR, receives two inputs: the aligned face of the target person, as extracted from I, and the aligned face in the generated frame f^0_i. In both cases, the face is extracted and aligned using the method of [2].

The face crop obtained from f^0_i is denoted c^0_i and serves as the input to FR. The face crop obtained from I and p^{M*} is denoted by c_I, and it serves as a conditioning signal to this network. For this purpose, the pre-trained VGGFace2 [3] network is used, and the activations of the penultimate layer, denoted by $\text{VGGFace}(c_I)$ are concatenated to the latent representation given by the encoder part of FR.

FR has the same autoencoder architecture as the de-id network [15], which solves the de-identification problem, which is very different from the current face refinement goal. We, therefore, employ a perceptual loss that differs from that of [15] and minimize the following loss:

$$L_{\text{facep}} = \sum_j \|\,\text{VGGFace}_j(c_I) - \text{VGGFace}_j(c^0_i)\| \tag{18}$$

where the index j is used to denote the spatial activations size at specific layers of network VGGFace, and the summation runs over the last layers of each block of size 112×112, 56×56, 28×28, 7×7, 1×1 (1×1 being the size of the topmost block, i.e., $\text{VGGFace}(c) = \text{VGGFace}_{1\times1}(c)$). The rest of the loss terms (reconstruction losses, mask regularization losses, adversarial losses) are the same as [15].

FR outputs a generated crop c and a blending mask m^c:

$$[c, m^c] = FR(c_I, c^0_i) \tag{19}$$

To create the final frame f_i, the crop c is blended with the region of frame f^0_i that corresponds to the face, in accordance with the values of the mask m^c.

4. Experiments

Datasets. Our networks are trained on cropped images, each containing a single person. The VIP dataset [55] is used to train the P2B network. The dataset contains 404 densely annotated videos with pixel-wise semantic part categories and a total of 21k frames. After cropping each separate person, the customized dataset contains a total of 62k images. The B2F network is trained by combining two datasets. MHPv2 [23] contains 25k images with an average of three people per image. After removing small and highly occluded people, 53k unique people remain. CIHP [17] contains 28k images. After pre-process, 1.7k different people with a total of 44k images (average of 25 images per person) remain. For each person, up to 15 random pairs are chosen, resulting in 19k unique pairs. Additional implementation details are provided in the supplementary.

For the numerical analysis, the target is taken from the driving video, establishing a valid ground-truth. For visual comparisons, where no ground-truth is required, we select 22 target images, out of which 12 are clearly visible, in a full-bodied frontal pose (denoted as the "simple" targets). Ten target images depict individuals who are not fully visible, or not in a standing frontal pose, denoted as the "challenging" targets. All target images used are provided in the supplementary. The vast majority of the selected target images are taken out of the DFDC dataset [8]. The DFDC dataset is uniquely diverse, allowing a comprehensive evaluation of the methods over different attributes, such as ethnicity, gender and age, but also pose, viewpoint and scale. Additional images were obtained from consenting individuals, attached as part of the supplementary.

Baselines. We compare our results with state of the art methods that represent the different approaches existing in the literature for the task of dance generation. When available, we use the authors' pre-trained weights; otherwise, we train the models with our dataset, following the authors' instructions. **fsV2V**[41] generates the entire video using a target image, OpenPose and DensePose data. It employs a hyper-network that predicts the weights of a vid2vid network. To achieve improved results, we followed the authors' instructions and fine-tuned the network for each

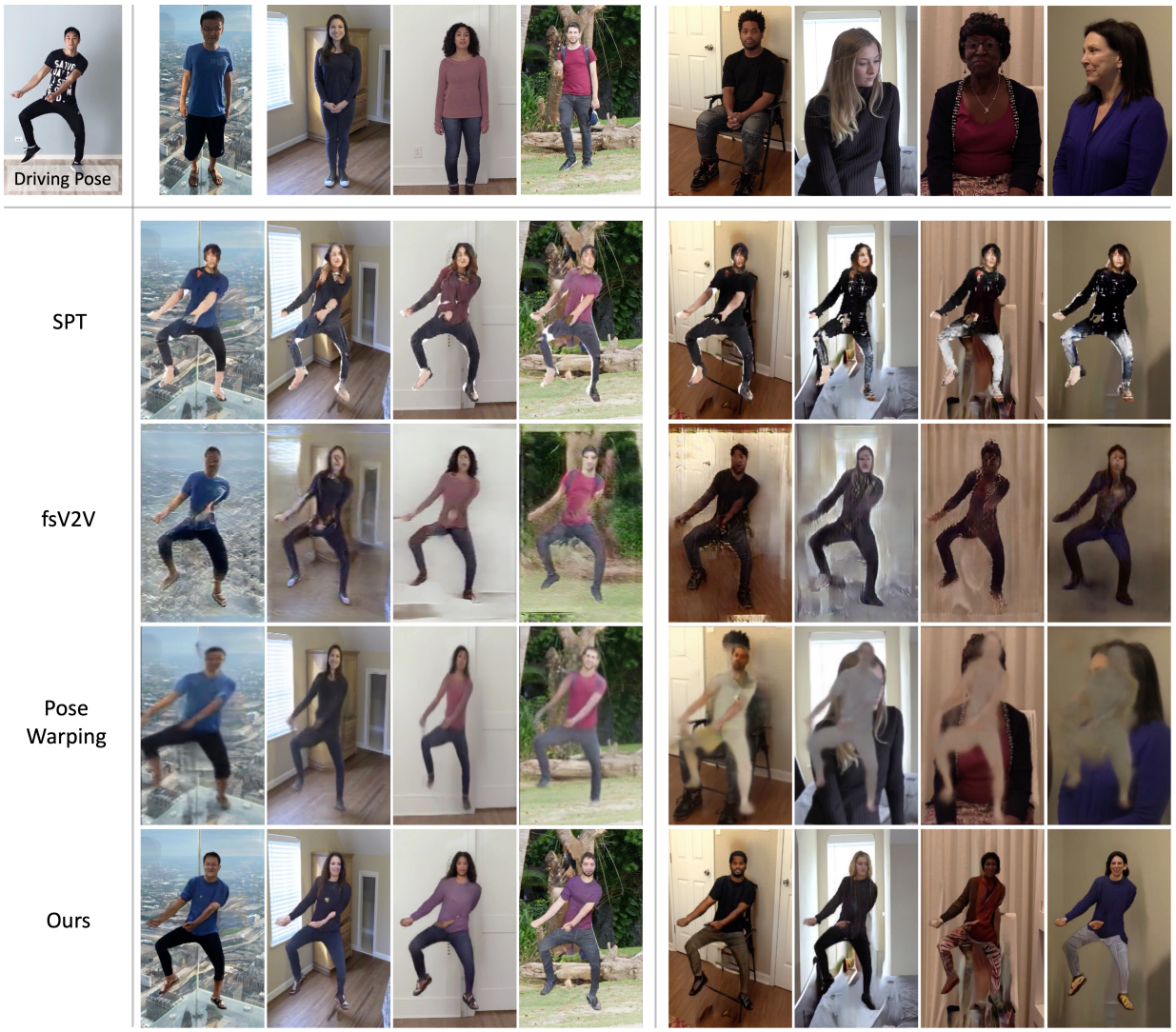


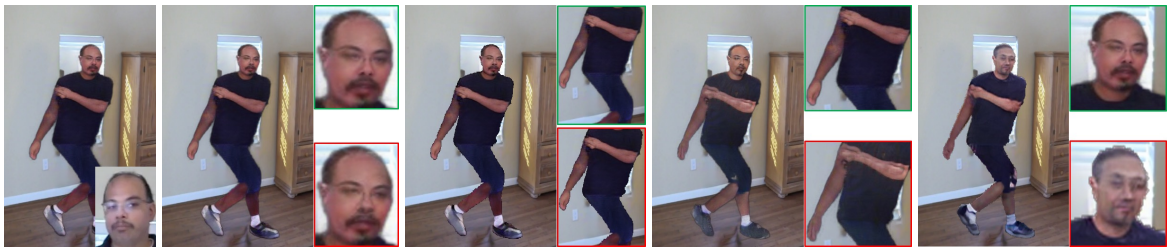
Figure 4. Comparison with previous methods. Each column presents a different target image. Our method is better able to handle both the "simple" (4 left) and "challenging" (4 right) targets, rendering higher quality and better appearance preserving results.

Method	SSBS ↑	SSIS ↑	DPBS ↑	DPIS ↑	LPIPS ↓ (VGG)	LPIPS ↓ (SqzNet)	SSIM ↓	FID ↓	Human Preference
fsV2V[41]	0.870	0.193	0.896	0.436	0.567	0.474	0.255	201.82	0.98
Pose Warping[1]	0.764	0.143	0.791	0.347	0.462	0.372	0.132	159.71	0.88
SPT[39]	0.851	0.165	0.862	0.404	0.378	0.289	0.127	109.13	0.81
Ours	**0.902**	**0.218**	**0.928**	**0.500**	**0.375**	**0.283**	**0.116**	**83.95**	-

Table 1. Comparison with previous work. The last column denotes the percent of samples in which the users preferred our results over the baseline. All results were obtained on "simple" targets only, as previous methods could not handle "challenging" targets.

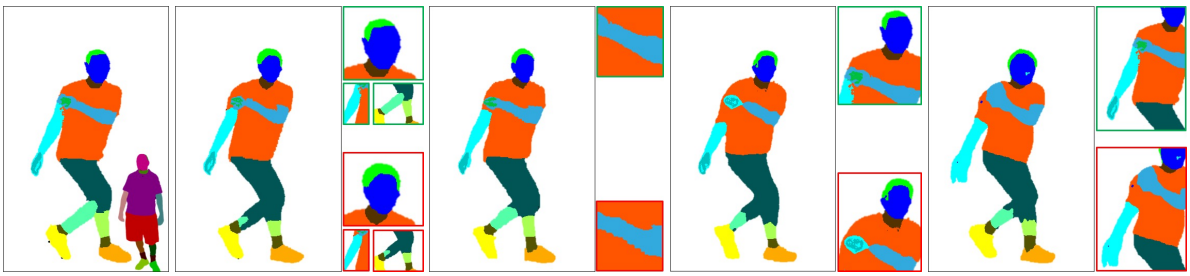
video. **Pose Warping**[1] generates a new frame by transforming each body part of the target, based on pose keypoints of the source and target images, followed by a fusion operation. **SPT**[39] resembles our approach, as the generator consists of two main parts. The first, a semantic generator, generates a new semantic map based on the source semantic segmentation and the new pose. The second, an appearance generator, renders the final frame. Generation is performed gradually in 128x128 and 256x256 pixels. Since the authors did not release the code for their semantic generator, we employ our P2B results instead.

Evaluation metrics. All comparisons are made over tar-



(a) (b) (c) (d) (e)

Figure 5. B2F/FR ablation study. (a) Our result and the target face. In the following, we show the resulting frame of a variant of our method. In red a zoom in of a certain part, and in green the same crop from our full method. (b) No FR (*blurrier face, features are less distinctive*), (c) no blending mask (*crude edges surrounding the entire rendered character*), (d) hand/finger labels not added (*arm distortions due to finger uncertainty, fingers less distinct*), (e) no face loss, lower resolution (256x160) (*appearance not preserved, edge pixelization*)



(a) (b) (c) (d) (e)

Figure 6. P2B ablations. (a) Our result and the target parsing (scaled down). The following are various variants. In red, a zoom in version, and in green the same zoom applied to the output of the full method. (b) No squeezing and stretching of the input/output parsing (*body structure, hair, and clothing less consistent*), (c) a less accurate version of DensePose is used (*boundary artifacts*), (d) DensePose is not used as input (*increased limbs artifacts, instability in body structure*), (e) no DP and no hand/finger labels (*enormous arms*).

gets and driving videos that do not appear in any training datasets. We use nine videos with an average of 300 frames each, obtained with consent from a video blogger. The evaluation metrics can be naturally divided into two distinct groups: quality and pose similarity.

For pose similarity, DPBS (DensePose Binary Similarity) and DPIS (DensePose Index Similarity) calculations [13] are used and are further adapted to serve as semantic segmentation similarity metrics (SSBS and SSIS). DPBS (SSBS) evaluates the IoU between a binary representation of the ground-truth and generated DensePose (the HP network), while DPIS (SSIS) evaluates the mean over each body-part index, for the same network.

For quality metrics, we rely on SSIM [44], LPIPS [51] and FID [19] to capture perceptual notions. LPIPS is applied with both the VGG [38] and SqueezeNet [20] networks. In addition, a user study is conducted among $n = 50$ participants. Each participant is shown the nine videos, where each video is shown as an instance generated by our method alongside an instance generated by one of the previous methods. The videos and targets are randomly selected such that three videos are presented for each method. The participant is asked to then select the video they prefer for each of the nine pairs of videos shown.

Results. Since the baseline methods struggle with challenging conditions, we measure performance only on the "simple" settings. As can be seen in Tab. 1, our method achieves superior results over all baselines and metrics. Those are apparent for both pose similarity and quality metrics. Additionally, the users present an overwhelming preference towards our method.

A visual comparison can be seen in Fig. 4 and in the supplementary (image and video samples). For both "simple" and "challenging" targets, our results are noticeably better at appearance preservation and quality.

4.1. Ablation

A visual ablation study is provided, where a distinction is made between structural and full pipeline aspects. The necessity of certain components in B2F and the existence of the FR network are examined with details in Fig. 5, while P2B is evaluated in Fig. 6. For each case, the dominant discrepancies are emphasized in a green square for our result, and a red square for each ablation case.

5. Conclusions

The desiderata of person animation techniques include not just visual quality, natural motion, motion fidelity, and appearance preservation, but also the ability to capture multiple body types, gender, ethnicity, and age groups. Diver-

sity in human pose generation is imperative to making sure technology is inclusive and can benefit everyone. However, it is often neglected in the literature.

The method we present, provides a much more detailed model of the human body, its appearance and its motion, than previous approaches. It is trained in a way that encourages it to address diverse inputs. In a comprehensive set of experiments, we demonstrate that the method is able to obtain better visual quality and better fidelity of both motion and appearance than the existing methods.

References

[1] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8340–8348, 2018. 2, 7

[2] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017. 3, 6

[3] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. *arXiv preprint arXiv:1710.08092*, 2017. 3, 6

[4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*, 2018. 3

[5] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. *arXiv preprint arXiv:1808.07371*, 2018. 2

[6] Patrick Chao, Alexander Li, and Gokul Swamy. Generative models for pose transfer. *arXiv preprint arXiv:1806.09070*, 2018. 2

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[8] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge dataset. *arXiv preprint arXiv:2006.07397*, 2020. 6

[9] Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. Soft-gated warping-gan for pose-guided person image synthesis. In *Advances in neural information processing systems*, pages 474–484, 2018. 2

[10] Patrick Esser, Johannes Haux, Timo Milbich, and Björn Ommer. Towards learning a realistic rendering of human behavior. In *ECCV WORKSHOP*, 2018. 2

[11] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8857–8866, 2018. 2

[12] Mikhail Figurnov, Aizhan Ibraimova, Dmitry P Vetrov, and Pushmeet Kohli. Perforatedcnns: Acceleration through elimination of redundant convolutions. In *Advances in Neural Information Processing Systems*, pages 947–955, 2016. 4

[13] Oran Gafni and Lior Wolf. Wish you were here: Context-aware human generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7840–7849, 2020. 8

[14] Oran Gafni, Lior Wolf, and Yaniv Taigman. Live face de-identification in video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9378–9387, 2019. 2

[15] Oran Gafni, Lior Wolf, and Yaniv Taigman. Live face de-identification in video. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 6

[16] Oran Gafni, Lior Wolf, and Yaniv Taigman. Vid2game: Controllable characters extracted from real-world videos. *arXiv preprint arXiv:1904.08379*, 2019. 2

[17] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 770–785, 2018. 5, 6

[18] David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016. 2

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017. 8

[20] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016. 8

[21] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 4

[22] Jianshu Li, Jian Zhao, Yunchao Wei, Congyan Lang, Yidong Li, Terence Sim, Shuicheng Yan, and Jiashi Feng. Multi-human parsing in the wild. *arXiv preprint arXiv:1705.07206*, 2017. 5

[23] Jianshu Li, Jian Zhao, Yunchao Wei, Congyan Lang, Yidong Li, Terence Sim, Shuicheng Yan, and Jiashi Feng. Multiple-human parsing in the wild. *arXiv preprint arXiv:1705.07206*, 2017. 6

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

[25] Jae Hyun Lim and Jong Chul Ye. Geometric gan. *arXiv preprint arXiv:1705.02894*, 2017. 6

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

[27] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou 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. 2

[28] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in Neural Information Processing Systems*, pages 406–416, 2017. 2

[29] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, 2017. 4

[30] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[31] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML)*, pages 807–814, 2010. 4

[32] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 5

[33] Albert Pumarola, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. Unsupervised person image synthesis in arbitrary poses. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[34] Iasonas Kokkinos Rĩza Alp Güler, Natalia Neverova. Densepose: Dense human pose estimation in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[35] Jian Ren, Menglei Chai, Sergey Tulyakov, Chen Fang, Xiaohui Shen, and Jianchao Yang. Human motion transfer from poses in the wild. *arXiv preprint arXiv:2004.03142*, 2020. 2

[36] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Advances in Neural Information Processing Systems*, pages 7135–7145, 2019. 2

[37] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3408–3416, 2018. 2

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3, 8

[39] Sijie Song, Wei Zhang, Jiaying Liu, and Tao Mei. Unsupervised person image generation with semantic parsing transformation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2, 7

[40] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 4

[41] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2, 6, 7

[42] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[43] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 4

[44] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 8

[45] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018. 2

[46] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Photo wake-up: 3d character animation from a single photo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5908–5917, 2019. 2

[47] Zhuoqian Yang, Wentao Zhu, Wayne Wu, Chen Qian, Qiang Zhou, Bolei Zhou, and Chen Change Loy. Transmomo: Invariance-driven unsupervised video motion retargeting. In *Computer Vision and Pattern Recognition*, 2020. 2

[48] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. *arXiv preprint arXiv:1806.03589*, 2018. 3

[49] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. *arXiv preprint arXiv:1801.07892*, 2018. 3

[50] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019. 6

[51] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 8

[52] Jian Zhao, Jianshu Li, Yu Cheng, Li Zhou, Terence Sim, Shuicheng Yan, and Jiashi Feng. Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing. *arXiv preprint arXiv:1804.03287*, 2018. 5

[53] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 2

[54] Qixian Zhou, Xiaodan Liang, Ke Gong, and Liang Lin. Adaptive temporal encoding network for video instance-level human parsing. In *Proc. of ACM International Conference on Multimedia (ACM MM)*, 2018. 4

[55] Qixian Zhou, Xiaodan Liang, Ke Gong, and Liang Lin. Adaptive temporal encoding network for video instance-level human parsing. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1527–1535, 2018. 4, 6

[56] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara Berg. Dance dance generation: Motion transfer for internet videos. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[57] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2347–2356, 2019. 2