## 1. Video Demonstration

Our video demo can be found at https://youtu.be/mSaIrz8lM1U and examples from our comparison to baselines and ablation study can be found at https://youtu.be/sQD0WVS0blg.

## 2. Implementation Details

Our generator and discriminator architectures are modified from pix2pixHD [41] to handle the temporal setting. We follow the progressive learning schedule from pix2pixHD and learn to synthesize at $512 \times 256$ at the first (global) stage, and then upsample to $1024 \times 512$ at the second (local) stage. For predicting face residuals, we use the global generator of pix2pixHD and a single $70 \times 70$ PatchGAN discriminator [16]. We set hyperparameters $\lambda_P = 5$ and $\lambda_{VGG} = 10$ during the global and local training stages respectively. For the dataset collected in Section **??**, we trained the global stage for 5 epochs, the local stage for 30 epochs, and the face GAN for 5 epochs.

For the perceptual loss $\mathcal{L}_P$, we compare the `conv1_1`, `conv2_1`, `conv3_1`, `conv4_1`, and `conv5_1` layer outputs of the VGG-19 network.

Our generator and discriminator architectures follow that presented by Wang et al. [41]. The fake-detector architectures matches that of the discriminator with a final fully connected layer.

## 3. Dataset Collection

Our dataset of long target videos consists of footage we filmed ourselves from 8 to 17 minutes with 4 videos at $1920 \times 1080$ resolution and 1 at $1280 \times 720$. Our goal in collecting a dataset of target videos is to provide the community with open-source data for which we explicitly collect release forms in which subjects allow their data to be released to other researchers. We recruited target subjects from different sources: friends, professional dancers, reporters etc. To learn the appearance of the target subject in many poses, it is important that the target video captures a sufficient range of motion and sharp frames with minimal blur. Similarly, we used a stationary camera to ensure a static background in all frames. To ensure the quality of the frames, we filmed our target subjects for between 8 and 30 minutes of real time footage at 120 frames per second using a modern cellphone camera, and use the first $20\%$ of the footage for training and the last $80\%$ for testing. Since our pose representation does not encode information about clothes and hair, we instructed our target subjects not to wear loose clothing and to tie up long hair.

In contrast, source videos can be easily collected online as we only require decent pose detections on these. We therefore use in-the-wild single-dancer videos where the only restriction we enforce is a static camera position.


Figure 1: We compare a lower resolution version of our model without a Face GAN (top) with a lower resolution of vid2vid [41]. We find our results comparable.

## 4. Comparison with vid2vid

We also compare our model with a concurrent video synthesis framework called vid2vid [**?**] as seen in Table **??**. The excessive requirement of memory and computing power of vid2vid prohibits us from comparing with their model in the high resolution setup. Instead, we train both our model and theirs in lower resolution ($512 \times 256$). Our system and vid2vid generally perform similarly and produce results of comparable quality. We provide a qualitative comparison in Figure 1.

## 5. Full Resolution Results

We include some examples of full resolution results in Figure 7.

## 6. Global Pose Normalization Details

In this section we describe our normalization method to match poses between the source and target. Consider a case where the source subject is significantly taller in frame than the target or is slightly elevated above the target subject's in frame position. If we directly input the unmodified poses to our system, we may generate images of the target person which are not congruent with the scene. In this example, the target person may appear large with respect to the background or surrounding objects, and may appear to be levitating since the input pose places the feet above the floor. Additionally, when generating an image from a very different pose from the in proportion and reasonably positioned poses in training, the overall quality of synthesis is expected to decline. Therefore we design a method to reasonably match the poses by finding a suitable transformation between the source and target poses. We parametrize this transformation in terms of a scale and translation factor applied to all pose keypoints for a given frame.

To find a suitable translation factor, we need to determine

Figure 2: Full resolution result on held out data.

Figure 3: Full resolution result on held out data.

the position of both subjects within their respective frames. We first find the closest position $s_{close}$ and farthest position $s_{far}$ the source subject is away from the camera in their video. Similarly, we do the same for the target by determining $t_{close}$ and $t_{far}$ respectively. The goal is then to map the close and far range of the source to that of the target subject

Figure 4: Full resolution result on held out data

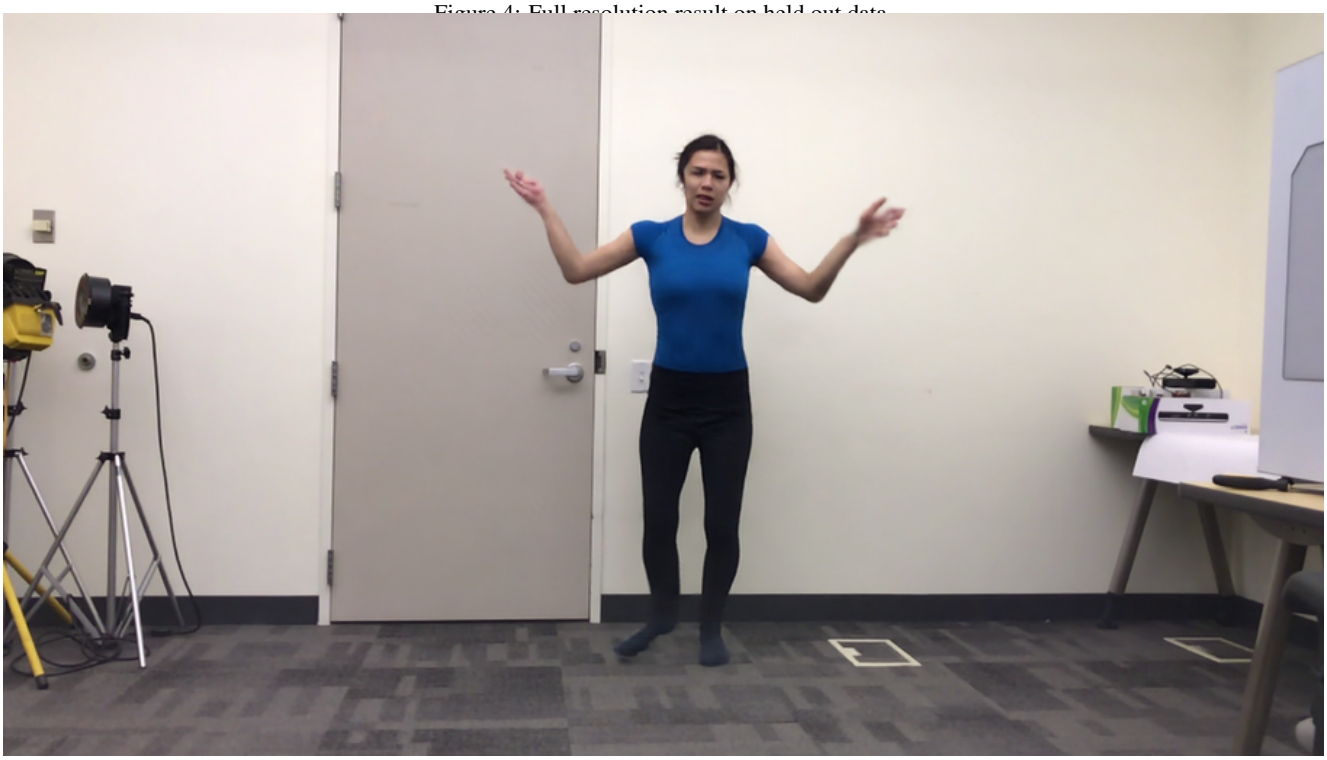

Figure 5: Full resolution result on held out data.

as to match the positions of both subjects, i.e. $s_{far} \mapsto t_{far}$ and $s_{close} \mapsto t_{close}$. Given a frame where the source is at position $y$, we then translate the source's pose vertically by:

$$translation = t_{far} + \frac{y - s_{far}}{s_{close} - s_{far}}(t_{close} - t_{far}) \quad (1)$$

Figure 6: Full resolution result on transfer (i.e. different source and target subjects).

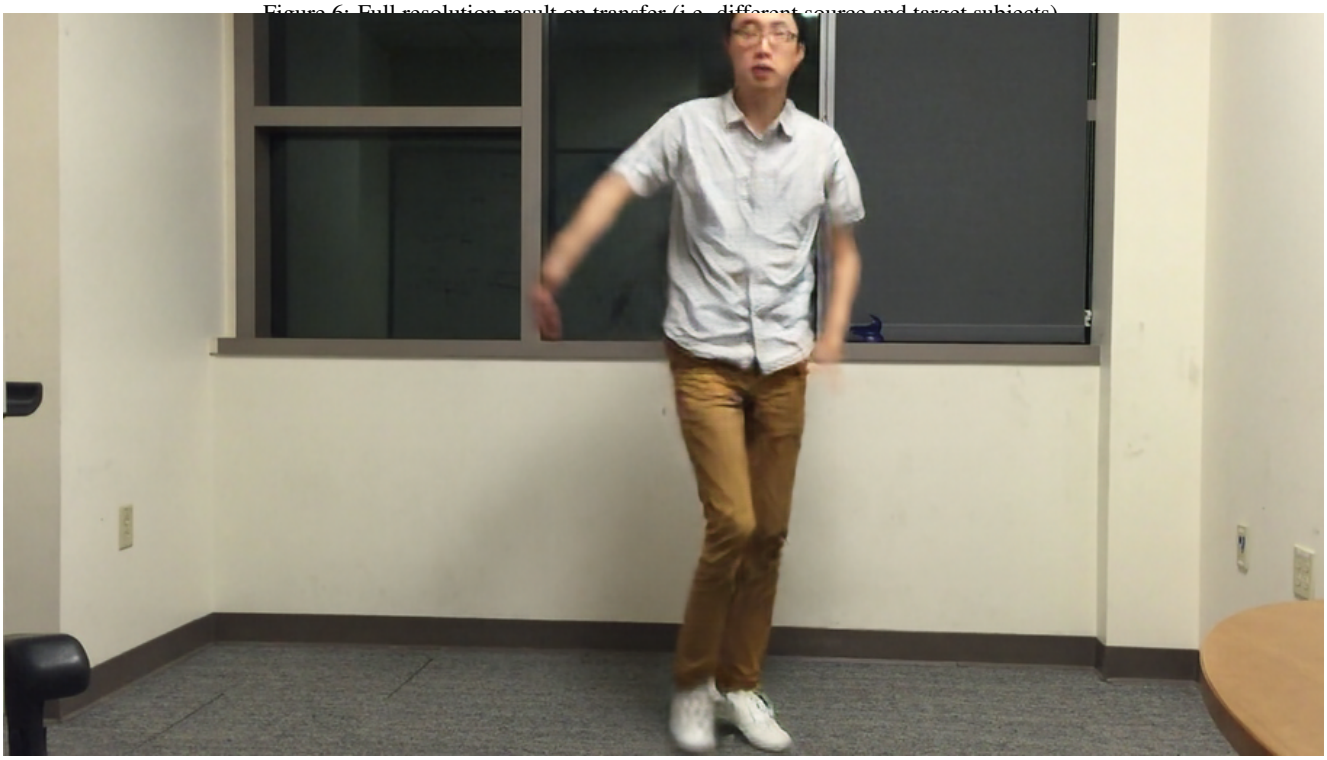

Figure 7: Full resolution result on transfer (i.e. different source and target subjects).

In practice, we use the average of the y coordinates of the subject's ankles to determine the position within a given frame.

To reasonably scale the source poses, we determine the heights of each subject at their closest and farthest positions in their video - denote these quantities as $h_{s_{close}}, h_{s_{far}}$ for

the source and $h_{t_{close}}, h_{t_{far}}$ for the target subjects respectively. We then determine separate scales for the close position given by $c_{close} = \frac{h_{t_{close}}}{h_{s_{close}}}$ and similarly for the far position given by $c_{far} = \frac{h_{t_{far}}}{h_{s_{far}}}$. When given a frame where the source is at position $y$, we scale the source's pose (in both x, y directions) by:

$$scale = c_{far} + \frac{y - s_{far}}{s_{close} - s_{far}}(c_{close} - c_{far}) \quad (2)$$

We use the euclidean distance between the average ankle position and the nose keypoint of our given pose as the subject's height in a given frame.

After the translation and scale factors have been determined for a given source pose, we then add the translation to all keypoints and then apply the scale factor so that the ankle $y$ positions remain the same (i.e. the ground is the x axis).

Given poses from a subject, we find the close position by taking the maximum y coordinate of their average ankle position over all frames.

$$s_{close} = \max\{\frac{s_{ankle1} + s_{ankle2}}{2}\}$$

The far position is found by clustering the y ankle coordinates which are less than (or spatially above) the median ankle position and about the same distance as the maximum ankle position's distance to the median ankle position. If we denote $S = \frac{s_{ankle1} + s_{ankle2}}{2}$ as the average ankle position in a given frame, then the clustering is as described by the set

$$\max\{S : ||S - s_{med}| < \alpha|s_{close} - s_{med}||\} \cap \{S < s_{med}\} \quad (3)$$

where $s_{med}$ is the median foot position, $max$ is the maximum ankle position, and $\epsilon$ and $\alpha$ are scalars. In practice we find setting $\alpha = 0.7$ generally works well, although this scalar can be finetuned on a case by case basis since it depends highly on the camera height and the subject's range of motion.