Do As I Do: Transferring Human Motion and Appearance between Monocular Videos with Spatial and Temporal Constraints

Thiago L. Gomes[1] Renato Martins[1,2] João Ferreira[1] Erickson R. Nascimento[1]
[1]Universidade Federal de Minas Gerais (UFMG), Brazil [2]INRIA, France
{thiagoluange,renato.martins,joaoferreira,erickson}@dcc.ufmg.br

Abstract

*Creating plausible virtual actors from images of real actors remains one of the key challenges in computer vision and computer graphics. Marker-less human motion estimation and shape modeling from images in the wild bring this challenge to the fore. Although the recent advances on view synthesis and image-to-image translation, currently available formulations are limited to transfer solely style and do not take into account the character's motion and shape, which are by nature intermingled to produce plausible human forms. In this paper, we propose a unifying formulation for transferring appearance and retargeting human motion from monocular videos that regards all these aspects. Our method synthesizes new videos of people in a different context where they were initially recorded. Differently from recent appearance transferring methods, our approach takes into account body shape, appearance, and motion constraints. The evaluation is performed with several experiments using publicly available real videos containing hard conditions. Our method is able to transfer both human motion and appearance outperforming state-of-the-art methods, while preserving specific features of the motion that must be maintained (*e.g.*, feet touching the floor, hands touching a particular object) and holding the best visual quality and appearance metrics such as Structural Similarity (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS).*

1. Introduction

Humans start learning early in their lives to recognize human forms and make sense of what emotions and meaning are being communicated by human movement. We are, by nature, specialists in the human form and movement analysis. Even for a meticulous artist, it may be hard to capture in a purely manual approach the fine details of human form and motion. Human form and motion estimation is at the core of a wide range of applications including entertainment, graphic animation, virtual and augmented reality, to name a few.

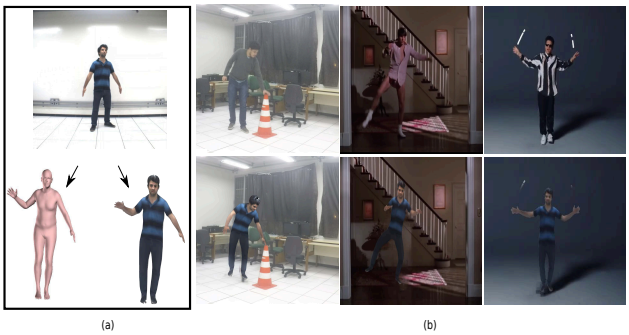

Figure 1: Overview of the motion and appearance transfer from a target video to different videos. After reconstructing a model for the target human (shown in (a)), we transfer his shape and motion to different videos as shown in (b). **Top row**: video with the source motion. **Bottom row**: New video with the retargeted motion and appearance of the target human model.

Capturing human geometry and motion has been improved over the decades through model-based and learning techniques. Computer Vision and Computer Graphics communities have progressively adopted learning techniques to automate the modeling and animation process of articulated characters. We have witnessed a variety of approaches used to extract articulated character patterns and capture three-dimensional motion, shape, and appearance [24, 17, 14, 10] from videos and still images from real actors. Despite remarkable advances in estimating 3D pose and shape, most of these methods only provide 3D meshes from the outer surfaces of objects, pose, and skeletons associated with those meshes. Even techniques such as the works of Chan *et al*. [10], Esser *et al*. [14], and Wang *et al*. [33] are limited to only transfer the appearance/style from one actor to another. In other words, these methods stretch or shrink the texture of a target actor to fit the texture in the movement instead of retargeting and fitting the motion into the actor (an alluring example is depicted in Figure 2).

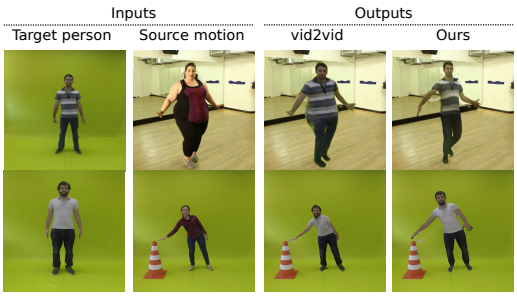

Figure 2: Motion and appearance transfer using vid2vid [33] and our formulation. From left to right: target person, source motion video with a human of different body shape, vid2vid, and our retargeting results. Note that vid2vid stretched, squeezed and shrinked the body forms whenever the transferring characters have different morphologies.

In this paper, we propose a novel retargeting framework that unifies appearance transfer with retargeting motion from video to video by adapting a motion from one character in a video to another character. The proposed approach synthesizes a new video of a person in a different context where this person was initially recorded. In other words, given two input videos, we investigate how to synthesize a new video, where a target person from the first video is placed into a new context performing different motions from the second video. The proposed method is composed of four main components: motion estimation in the source video, body model reconstruction from target video, motion retargeting with spatio-temporal constraints, and finally image composition. By imposing spatial and temporal constraints on the joints of the characters, our method preserves features of the motion, such as feet touching the floor and hands touching a particular object. Also, our method employs an adversarial learning in the texture domain to improve textures extracted from frames and leverage details in the visual appearance of the target person.

In this context, several recent learning-based methods have been proposed on synthesizing new pose from a source image (*e.g.*,[19, 40, 22, 10, 14]). Unlike the methods [10, 14, 33] that are built on learning approaches to work in the image domain to transfer texture, our approach aims at adapting the movement from one actor to another taking into account the main factors for a moving actor: body shape, appearance and motion.

The main technical contributions of this paper are as follows: i) a marker-less human motion estimation technique that takes into account both body shape and camera pose consistencies along the video; ii) a generative adversarial network for improving visual details that works directly with texture maps to restore facial texture of human models; and iii) a unified methodology carefully designed to transfer motion and appearance from video to video that preserves the main features of the human movement and retains the visual appearance of the target character.

We demonstrate the effectiveness of our approach quantitatively and qualitatively using publicly available video sequences containing challenging problem conditions, as shown in Figure 1.

2. Related Work

3D human shape and pose estimation. Several works have been proposed to estimate both the human skeleton and 3D body shape from images. Sigal *et al*. [28] compute shape by fitting a generative model (SCAPE [6]) to the image silhouettes. Bogo *et al*. [8] proposed the SMPLify method, which is a fully automated approach for estimating 3D body shape and pose from 2D joints in images. SMPLify uses a CNN to estimate 2D joint locations and then fits an SMPL body model [21] to these joints. Lassner *et al*. [20] take the curated results from SMPLify to train 91 keypoint detectors. Some of these detectors correspond to the traditional body joints, and others correspond to locations on the surface of the body. Similarly, Kanazawa *et al*. [17] used unpaired 2D keypoint annotations and 3D scans to train an end-to-end network to infer the 3D mesh parameters and the camera pose. Their method outperformed the works [8, 20] regarding 3D joint error and runtime. However, their bounding box cropping strategy, which frees 3D pose regression from having to localize the person in scale and image space, loses global information and temporal consistency required in the motion transfer.

Retargeting motion. Gleicher seminal work of retargeting motion [15] addressed the problem of transferring motion from one virtual actor to another with different morphologies. Choi and Ko [11] pushed further Gleicher's method by presenting an online version based on inverse rate control. Villegas *et al*. [31] proposed a kinematic neural network with an adversarial cycle consistency to remove the manual step of detecting the motion constraints. In the same direction, the recent work of Peng *et al*. [24] takes a step towards automatically transferring motion between humans and virtual humanoids. Despite remarkable results in transferring different movements, these methods are limited to either virtual or textureless characters. Similarly, Aberman *et al*. [2] proposed a 2D motion retargeting using a high-level latent motion representation. This method has the benefit of not explicitly reconstructing 3D poses and camera parameters, but it fails to transfer motions if the character walks towards the camera or with a large variation of the camera's point-of-view.

Synthesizing views. The past five years has witnessed the explosion of generative adversarial networks (GANs)

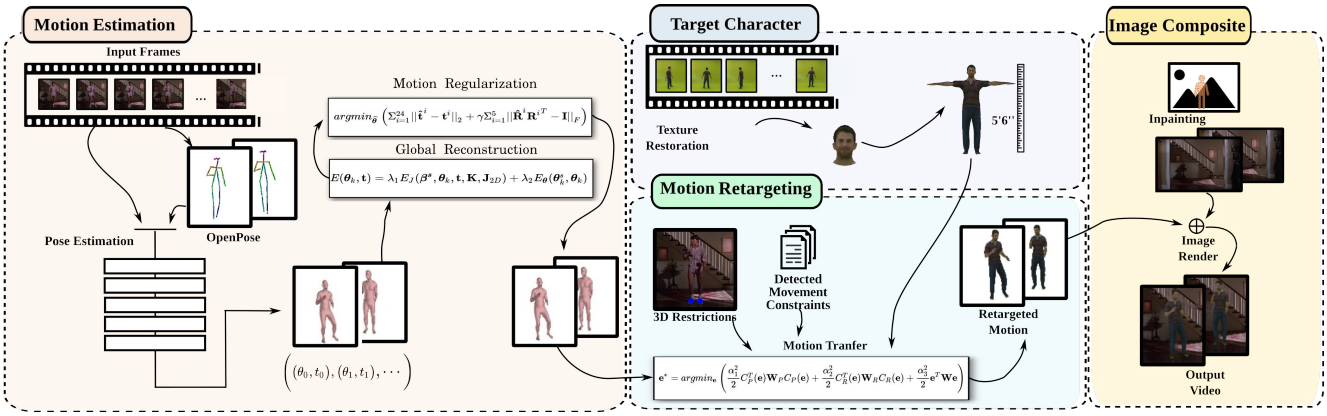

Figure 3: Overview of our retargeting approach that is composed of four main components: human motion estimation in the source video (first component); we retarget this motion into a different target character model (second component), considering the motion constraints (third component) and by last, we synthesize the appearance of the target character into the source video.

to new view synthesis. GANs have emerged as promising and effective approaches to deal with the tasks of synthesizing new views, against image-based rendering approaches (*e.g.*,[18, 38, 27]). More recently, the synthesis of views is formulated as being a learning problem (*e.g.*, [30, 13, 36, 7, 14]), where a distribution is estimated to sample the new views. A representative approach is the work of Ma *et al*. [22], where the authors proposed to transfer the appearance of a person to a given pose in two steps. Similarly, Lassner *et al*. [19] proposed a GAN called ClothNet. ClothNet produces people with similar pose and shape in different clothing styles given a synthetic image silhouette of a projected 3D body model. In the work of Esser *et al*. [14], a conditional U-Net is used to synthesize new images based on estimated edges and body joint locations. Despite the impressive results for several inputs, in most cases, these methods fail to synthesize details of the human body such as face and hands. Recent works [1, 10] applied an adversarial training to map a 2D source pose to the appearance of a target subject. Although these works employ a scale-and-translate step to handle the difference in the limb proportions between the source skeleton and the target, they have still clear gaps in the motion in the test time when comparing with the motion in the training time. Wang *et al*. [33] presented a general video-to-video synthesis framework based on conditional GANs to generate high-resolution and temporally consistent videos. Unfortunately, these learning-based techniques transfer style and wrongly distorts characters with different morphologies (proportions or body parts' lengths). Moreover, differently from our method, these state-of-the-art approaches [1, 10, 33] are dataset specific, *i.e.*, they require training a different GAN for each video of the target person with different motions to perform the transferring. This training is computationally intensive and takes several days on a single GPU. Our method, for its turn, does not require a large number of images and powerful hardware for training, keeps visual details from the target character while preserving the features of the transferred motion.

3. Retargeting Approach

Our method can be divided into four main components. We first estimate the motion of the character in the source video. Our *motion estimation* regards essential aspects to obtain plausible character movements, such as of ensuring a common system coordinate for all image frames and temporal motion smoothness. Second, we extract the *body shape and texture* of the target character in the second video. Then, the *retargeting* component adapts the estimated movement to the body shape of the target character, while considering temporal motion consistency and the physical interactions (constraints) with the environment. Finally, the *image rendering and composition* component renders the texture (appearance), extracted from the target character, into the background of the source video. Figure 3 shows a schematic representation of the method pipeline.

3.1. Human Body and Motion Representation

We represent the human motion by a set of translations and rotations over time of joints that specify a human skeleton. This skeleton is attached to the characters body and is defined as a hierarchy of 24 linked joints. Each joint pose $\mathbf{P}^i$ ($\mathbf{P}^i \in \mathbb{SE}(3)$ is the pose of the i-th joint) is given by recursively rotating the joints of the skeleton tree, starting from the root joint and ending in its leaf joints (*i.e.*, the forward kinematics denoted as FK). To represent the 3D shape of the human body, we adopted the SMPL model parametrization [21], which is composed of a learned human shape distribution $\mathcal{M}$, 3D joint angles ($\boldsymbol{\theta} \in \mathbb{R}^{72}$ defining 3D rotations of the skeleton joint tree), and shape coef-

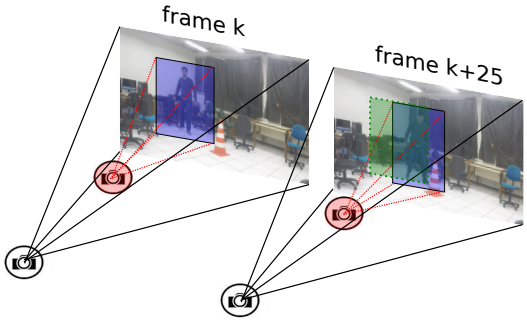

Figure 4: Schematic view of the motion reconstruction. Note the change of the point of view between the virtual cameras (in red) for a gap of 25 frames (showed by the different positions of the blue and green boxes). This change is ignored by the bounding box crop, producing temporally inconsistent pose and shape estimates.

ficients $\boldsymbol{\beta} \in \mathbb{R}^{10}$ that model the proportions and dimensions of the human body.

3.2. Human Motion Estimation

We start estimating the actor's motion in the source video. Our method builds upon the learning-based SMPL pose estimation framework of Kanazawa *et al*. [17]. The human pose and shape are predicted in the coordinate system of a bounding box around the person, where a weak-perspective camera model is adopted as shown in Figure 4. This bounding box normalizes the person in size and position, as also noted in [23], which frees 3D pose estimation from the burden of computing the scale factor (between the body shape to the camera distance) and the location in the image. However, this incurs in a loss of temporal pose consistency required in the motion transfer. This also often leads to wrong body shape estimates for each frame, which should be constant along the video.

In order to overcome these issues, we map the initial pose estimation using virtual camera coordinates, as illustrated in Figure 4. For that, our motion estimation minimizes an energy function with two terms:

$$E(\boldsymbol{\theta}_k, \mathbf{t}) = \lambda_1 E_J(\boldsymbol{\beta^s}, \boldsymbol{\theta}_k, \mathbf{t}, \mathbf{K}, \mathbf{J}_{2D}) + \lambda_2 E_{\boldsymbol{\theta}}(\boldsymbol{\theta}^s_k, \boldsymbol{\theta}_k), \tag{1}$$

where $\mathbf{t} \in \mathbb{R}^3$ is the translation, $\mathbf{K} \in \mathbb{R}^{3\times3}$ is the camera intrinsic matrix, $\mathbf{J}_{2D}$ is the projections of the joints in the reconstruction of [17], and λ_1, λ_2 are scaling weights. The first term encourages the pose projections of the joints to remain in the same locations into the common reference coordinate system. The second term favors maintaining the joints' angles configuration, while reinforcing the adopted character shape to have averaged shape coefficients ($\boldsymbol{\beta^s}$) of the entire video. Finally, the human model pose in each frame is then obtained with the forward kinematics (FK) in the skeleton tree:

$$\left(\mathbf{P}^0_k\ \mathbf{P}^1_k\ \ldots\ \mathbf{P}^{23}_k\right) = \text{FK}(\mathcal{M}, \boldsymbol{\beta^s}, \boldsymbol{\theta_k}), \tag{2}$$

where $\mathbf{P}^k_i = [\text{FK}(\mathcal{M}, \boldsymbol{\beta}^s, \boldsymbol{\theta}^s_k)]_i$ is the pose of the joint i^{th} at frame k. Thus, we define the motion of each joint i as the set of successive poses in the frames $\mathbf{M}^i = [\mathbf{P}^i_1\ \mathbf{P}^i_2\ ...\ \mathbf{P}^i_n]$.

Motion Regularization. Since the character poses are estimated frame-by-frame, the resulting motion might present some shaking motion with high-frequency artifacts in some short sequences of the video. To reduce these effects, we apply a motion reconstruction to seek a new set of joint angles $\widehat{\boldsymbol{\theta}^s}$ that creates a smoother character motion. We compute a smoother configuration for the joints by minimizing the following cost of inverse kinematics (IK) in the joint positions and end-effectors orientations $\mathbf{M}^i$:

$$\underset{\widehat{\boldsymbol{\theta}}}{\text{argmin}} \left(\Sigma^{24}_{i=1}||\hat{\mathbf{t}}^i - \mathbf{t}^i||_2 + \gamma\Sigma^5_{i=1}||\hat{\mathbf{R}}^i{\mathbf{R}^i}^T - \mathbf{I}||_F\right), \tag{3}$$

where $\widehat{\mathbf{P}}^i = [\hat{\mathbf{R}}^i\ \hat{\mathbf{t}}^i]$ is given by the forward kinematics $\mathbf{P}^i_k = [\text{FK}(\mathcal{M}, \boldsymbol{\beta}^s, \boldsymbol{\theta}^s_k)]_i$ with unknown joint angles $\boldsymbol{\theta}^s_k$, $||.||_F$ is the Frobenius norm of the orientation error, and γ the scaling factor between the position of all joints and orientation of the end-effectors (*i.e*., feet, hands, and head). This reconstruction strategy removes high-frequency artifacts of the motion while maintaining the main movement features of the body end-effectors.

3.3. Target 3D Human Body Model Building

This section presents our strategy to build the 3D model and texture of the character that is transferred to the source video (*i.e*., the target body model $\boldsymbol{\beta}^t$). Our target reconstruction component starts with an initial 3D body model from Alldieck *et al*. [5]. This produces a reasonable model of people in clothing from a single video in which the person is moving in an A-pose configuration. We remark that any technique, capable of creating plausible 3D human models, could be used to get this initial body model estimate in our method (*e.g*., [4, 3, 26]). Although the good resulting 3D human model accuracy, the texture images were often blurred and lacking of details. In the following, we discuss how to mitigate the loss of detail by taking inspiration from the recent advance in generative adversarial networks.

GAN Face Texture Restoration. According to Balakrishnan *et al*. [7], humans are particularly good at detecting facial abnormalities such as deformations or blurring. Unfortunately, when mapping textures into a target 3D model, we lose important details mainly because of warping and interpolation artifacts.

In order to reduce this effect, we exploit the capability of generative adversarial networks (GANs) to denoise images [16]. However, differently from previous works,

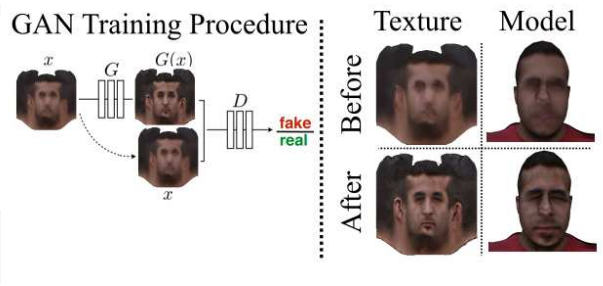

Figure 5: Face texture enhancement strategy using a conditional GAN. An example of the restoring results in the texture domain is shown in the right column, side by side with the visualizations of the textures in the human model.

we perform the learning directly in texture map images as shown in Figure 5. This produced better restoration results probably due to smaller geometrical variability from the texture maps compared to the appearance in the 3D body mesh. We circumvent the problem of nonexistence of a publicly available dataset of face textures, to train our GAN model, by using 500 human face textures (real textures) from 3D.SK[1]. We augmented the training dataset by adding noise, small rotations and blurring warped images. For the training, we adopted the conditional GAN proposed in [16] used for image-to-image translation. Some input images of the augmented dataset and the resulting restoration can be seen in Figure 5 and in the supplementary material.

3.4. Retargeting using Space-time Constraints

After computing the source motion ($\mathbf{M}^i$, $\boldsymbol{\theta}^s$) and the target human 3D model ($\boldsymbol{\beta}^t$), we proceed to the motion retargeting component. The retargeting is essential to guarantee that some physical restrictions are still valid during the target character animation. In this paper, we assume that the target character has a homeomorphic skeleton structure to the source character, *i.e.*, the main geometric differences are in terms of bone lengths or proportions. Our first goal is to retain the joint configuration of the target as close as possible of the source joint configurations, *i.e.*, to keep the pose error $\mathbf{e}_k = \boldsymbol{\theta}^t_k - \boldsymbol{\theta}^s_k$ small and then preserve the appearance of the motion whilst respecting movement constraints. A secondary objective is to keep a similar *movement style* in the retargeted motion over time. Thus, we propose a prediction error in 3D space to maintain the style from the original character motion:

$$C_P = \sum_{k=i}^{i+n} \big(\text{FK}(\mathcal{M}, \boldsymbol{\beta}^t, \boldsymbol{\theta}^t_{k+1}) - \text{FK}(\mathcal{M}, \boldsymbol{\beta}^t, \boldsymbol{\theta}^t_k) - \big(\text{FK}(\mathcal{M}, \boldsymbol{\beta}^s, \boldsymbol{\theta}^s_{k+1}) - \text{FK}(\mathcal{M}, \boldsymbol{\beta}^s, \boldsymbol{\theta}^s_k)\big)\big)\,. \quad (4)$$

[1]https://www.3d.sk/

Rather than considering a full horizon cost (total number frames), we leverage only the frames belonging to a neighboring temporal window of n frames equivalent to two seconds of video. This neighboring temporal window scheme allows us to track the local temporal motion style producing a motion that tends to be natural compared with a realistic looking of the estimated source motion. Only considering a local neighboring window of frames also results in a more efficient optimization.

Spatial Motion Restrictions and Physical Interactions. The motion constraints are used to identify key features of the original motion that must be present in the retargeted motion. The specification of these constraints typically involves only a small amount of work in comparison with the task of creating new motions. Typical constraints are, for instance, that the target character feet should be on the floor; holding hands while dancing or while grabbing/manipulating an object in the source video. Some examples of constraints are shown in Figures 6 and 7, where the characters are placing their left hand in a box or over a cone object.

Our method is capable of adapting to such situations in terms of position by constraining the positioning of the end-effectors to respect a set of constraints in the frame k given by the joint poses $\mathbf{P}_R = \left(\mathbf{P}^j\ \mathbf{P}^m\ \ldots\ \mathbf{P}^n\right)$ as:

$$C_R = \Sigma_i \left([\text{FK}(\mathcal{M}, \boldsymbol{\beta}^t, \boldsymbol{\theta}^t_k)]_i - [\mathbf{P}_R]_i\right). \quad (5)$$

Space-time Cost Error Optimization. The final motion retargeting cost combines the source motion *appearance* with the different shape and restrictions of the target character using equations (4) and (5):

$$\mathbf{e}^* = \underset{\mathbf{e}}{\text{argmin}} \left(\frac{\alpha_1^2}{2} C_P^T(\mathbf{e}) \mathbf{W}_P C_P(\mathbf{e}) + \frac{\alpha_2^2}{2} C_R^T(\mathbf{e}) \mathbf{W}_R C_R(\mathbf{e}) + \frac{\alpha_3^2}{2} \mathbf{e}^T \mathbf{W} \mathbf{e} \right), \quad (6)$$

where $\mathbf{e} = (\mathbf{e}_{k+1}, \ldots, \mathbf{e}_{k+n})^T$, n the number of frames considered in the retargeting, α_1, α_2 e α_3 are the contributions for the different error terms and, $\mathbf{W}_P$, $\mathbf{W}_R$ and $\mathbf{W}$ are diagonal matrices of weights for the prediction, restrictions and motion similarity terms. Each of these weight matrices are set such as to penalize more the errors in joints that are closer to the root joint. We minimize this cost function with a gradient-based NLSQ iterative optimization scheme, where the Jacobians are computed using automatic differentiation for each degree of freedom. The optimization stops when either the error tolerance or the maximum number of iterations are reached. An example of the retarget motion

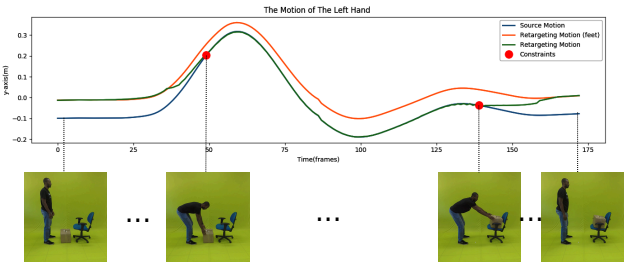

Figure 6: The left hand's trajectory on the y-axis when transferring the motion of *pick up a box* between two differently sized characters: original motion (blue line), a naive transfer without constraints at the person's hand (red line) and with constraints (green line). Frames containing motion constraints are located between the red circles.

trajectory of the left hand of our approach is shown in Figure 6. Note the smooth motion adaptation produced by the retargeting with the restrictions in frames 47 and 138 (green line) when the character's hand was touching the box.

3.5. Model Rendering and Image Compositing

The last step of your framework composed the rendered target character and the source background. For that, we first segment the source image into a background layer using, as a mask, the projection of our computed model with a dilation. Next, the background is filled with the method proposed by Criminisi *et al*. [12] to ensure temporal smooth to the final inpainting. We compute the final pixel color value as the median value between the neighboring frames. Finally, the background and our render of the model are composed using Poisson blender [25] to illumination adjustment. We remark that we tested different inpainting formulations, comprising deep learning-based methods presented in [37, 32]. Our experiments showed that although these deep learning-based methods synthesize plausible pixels for each frame, the adopted inpainting strategy has better results considering the visual and spatio-temporal consistency between the frames. Furthermore, [32] requires a static mask in the video and this is too restrictive to our problem.

4. Experiments and Results

Video Sequences. The selected videos cover a variety of representative conditions to the problem, such as different types of motion, lighting conditions and background, actors morphologies and videos used by previous works. All four sequences contain different subjects and types of motion constraints that should be taken into account in order to synthesize plausible videos. A short description of these videos is as follows: i) **alfonso-ribeiro**[2]: This video has strong illumination changes and fast character motions, with a 1.67 meters height male character dancing. The restrictions are mostly in the dancers' feet; ii) **joao-pedro**: Video with moderate speed motions and with a static background, where a 1.80 meters height male character is walking and interacting with a cone in the floor. The motion constraints are in the feet and hands; iii) **tom-cruise**[3]: This video contains motions with moderate speed but with displacements in all directions in the scene, where a 1.70 meters height male character is pretending to sing while dancing. The motion restrictions are in the dancer's feet; iv) **bruno-mars**[4]: Video with fast motions where a 1.65 meters height male character is dancing, with partial occlusions of arms and feet. The restrictions are in the dancer's feet. This sequence was also used by [10].

In order to build the target person shape and appearance, we used videos of People-Snapshot dataset [5] and videos of actors recorded with a Point Grey camera. These sequences were captured with the camera fixed at a distance of three meters from the characters.

Parameters Setting and GAN Training. We set $\lambda_1 = 10^{-6}$ and $\lambda_2 = 10^{-2}$ in the motion estimation. In the reconstruction and retargeting steps, we used $\gamma = 10$, $\alpha_1 = 10$, $\alpha_2 = 5$ and $\alpha_3 = 1$. Our textured dataset augmentation process was performed applying random small pixel translations (between -15 and 15 pixels) and random rotations (between -25 and 25 degrees) for the same image. Each original texture map was replicated twenty times with random transformations resulting in a training set of $10,000$ images. As suggested by Isola *et al*. [16], our training loss is a traditional GAN loss combined with a $L1$ loss to reduce including visual artifacts. We used a factor $\lambda = 500$ (conversely to $\lambda = 100$ employed in [16]) in order to avoid including visual artifacts. The other remaining training parameters were the Adam solver with learning rate of 0.0002 and momentum $\beta_1 = 0.5$, $\beta_2 = 0.999$.

Baseline and Metrics. We used the V-Unet proposed by Esser *et al*. [14] as a baseline. The baseline choice follows two main reasons: i) Most of the related work to our approach are in the area of image-to-image translation using conditional GANs and the V-Unet is a recent state-of-the-art technique that represents this class of approaches; ii) Recent state-of-the-art techniques such as [1, 10, 33] are dataset specific, *i.e*., they need to train a GAN for each video where the target subject is performing a large set of different poses. This training has a consequent computational effort, which can last several days. Furthermore, they did not provide the code or training data making the comparison impractical. Lastly, we recall that these methods are limited to transfer solely style and suffers from the structure issue previously discussed in Section 2 and shown in Figure 2.

[2]https://www.youtube.com/watch?v=pbSCWgZQf_g

[3]https://www.youtube.com/watch?v=IUj79ScZJTo

[4]https://www.youtube.com/watch?v=PMivT7MJ41M

Table 1: Visual quantitative metrics of our method and V-Unet.

Video sequence	SSIM[1]		LPIPS[2]		Missed detections[3]	
	V-Unet	Ours	V-Unet	Ours	V-Unet	Ours
alfonso-ribeiro	0.834	**0.837**	0.137	**0.126**	0.554	**0.342**
joao-pedro	0.980	**0.987**	0.018	**0.009**	0.596	**0.513**
tom-cruise	0.986	**0.988**	0.013	**0.008**	0.867	**0.832**
bruno-mars	0.950	**0.962**	0.044	**0.035**	**0.245**	0.301

[1] *Better closer to 1.* [2] *Better closer to 0.* [3] *Better closer to 0.*

Due to the lack of ground truth data for retargeting between two different video subjects, we adopted the same quantitative visual evaluation metrics of [10]. Thus, we measure the appearance quality using Structural Similarity (SSIM) [34] and Learned Perceptual Image Patch Similarity (LPIPS) [39] between consecutive frames. For a fair comparison, the final image composition of the V-Unet uses the same background inpainting and post-process used in our method. We also report the average number of missed 2D joints' detections from OpenPose [9, 29, 35] in the produced videos. Table 1 shows the quantitative appearance metrics and Figure 8 depicts some frames for all video sequences.

4.1. Discussion

Visual Appearance Analysis. It can be seen from the quantitative and qualitative results for all sequences that the proposed method leads the performance in both SSIM and LPIPS metrics. Furthermore, Figure 8 shows that our method presents a much richer and detailed visual appearance of the target character than when using V-Unet. One can easily recognize the person from the target video (top left image) in the samples of the retargeting video (third rows for each video).

To assets in which extent our texture denoising contributes to the success of our approach in retaining the visual appearance of the target, we tested our network restoration in several face textures from People-Snapshot dataset [5] and in our generated target human models. Figure 5 shows some results after applying the denoising GAN to improve the texture for typical faces. We provide some additional denoising results in the supplementary material.

Shape and Motion analysis. We show in Figure 8 some resulting frames for all four video sequences. Since V-Unet uses a global pose normalization, it resizes the source image to approximate scale and location of the target person and, then, it was not able to maintain the length of the limbs during the transferring. As a result, the limbs were stretched to fit the source shape. Conversely, the proposed approach did not stretch or shrink the body forms because it regards shape, appearance as well as the motion constraints to define the form of the retarget character.

In terms of motion reconstruction, our method also outperformed V-Unet. For instance, V-Unet clearly failed to place the target's feet on the right position in the last frame of alfonso-ribeiro results shown in Figure 8. Due to the temporal reconstruction, our method was able to transfer the target to the correct pose. Additionally, these results reinforce the capability of our method to impose different space-time constraints to the retargeting motion. As shown on the frames of Figure 8, different motions are adapted to fit the proportions of the target person and to keep the constraints of the motion, such as of placing the hand on cone object, as illustrated in the results from the sequence joao-pedro in Figures 7 and 8 for two distinct target characters. We provide additional results in the supplementary material.

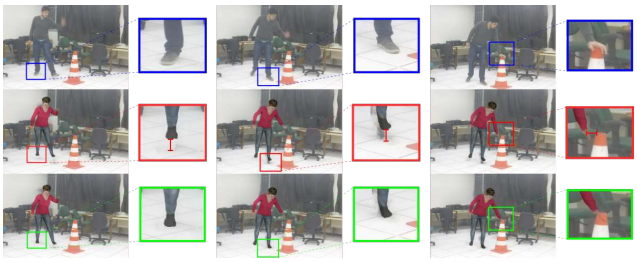

Figure 7: Video sequence results with motion constraints. **Top row:** source video. **Middle row**: results using a naive transferring without retargeting constraints. **Bottom row**: obtained results with our method considering the motion retargeting constraints.

5. Conclusions

In this paper, we proposed a complete retargeting framework that incorporates different strategies to extract and to transfer human motion, shape, and appearance between two real characters in monocular videos. Differently from classic retargeting methods that use either appearance or motion information, the proposed framework takes into account simultaneously four important factors to retargeting, *i.e.*, pose, shape, appearance, and features of the motion.

We performed real transferring experiments on publicly available videos. Our approach outperforms V-Unet in terms of both appearance metrics (SSIM and LPIPS) and number of missed joints' detections when estimating the skeleton. Our results suggest that retarget strategies based on image-to-image translation are not powerful enough to retarget motions while keeping the desired constraints of the motion and shape/appearance. Future work directions include automatically detecting the retargeting motion constraints in the videos, as well as improving the appearance restoration and transferring beyond the face texture maps. Another interesting topic would be to improve the scene compositing (*e.g.*, estimating the scene illumination) for a more realistic rendering.

Acknowledgments. The authors would like to thank CAPES, CNPq, and FAPEMIG for funding different parts of this work. We also thank NVIDIA Corporation for the donation of a Titan XP GPU used in this research.

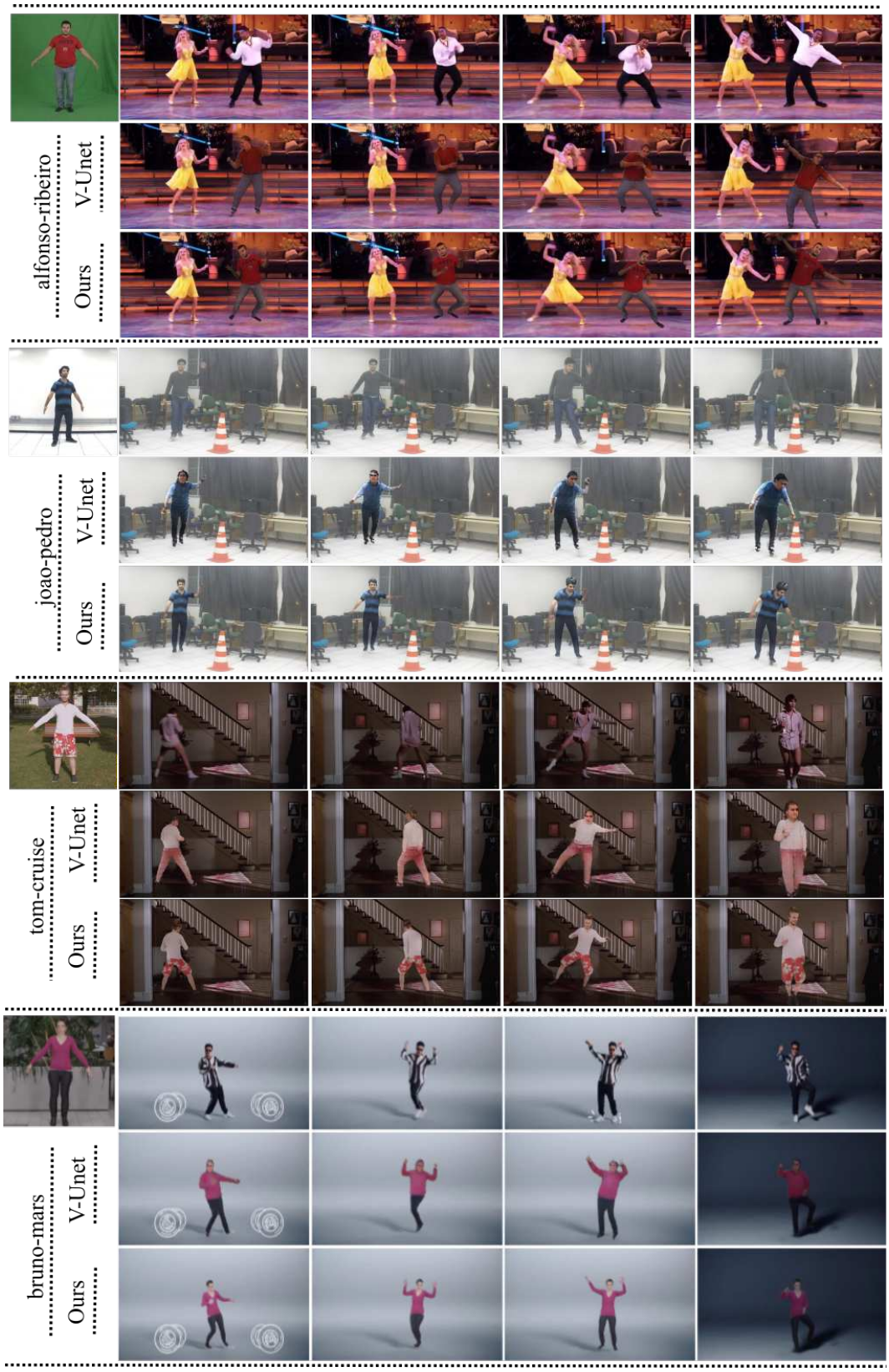

Figure 8: Qualitative retargeting results using video sequences with different types of motion, lighting conditions, background and actors morphologies. In each sequence: First row: target person and motion source; Second row: V-Unet result; Third row: Our method.

References

[1] K. Aberman, M. Shi, J. Liao, D. Lischinski, B. Chen, and D. Cohen-Or. Deep video-based performance cloning. *CoRR*, abs/1808.06847, 2018. 3, 6

[2] K. Aberman, R. Wu, D. Lischinski, B. Chen, and D. Cohen-Or. Learning character-agnostic motion for motion retargeting in 2d. *ACM Transactions on Graphics (TOG)*, 38(4):75, 2019. 2

[3] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll. Learning to reconstruct people in clothing from a single RGB camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2019. 4

[4] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Detailed human avatars from monocular video. In *International Conference on 3D Vision*, pages 98–109, Sep 2018. 4

[5] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 4, 6, 7

[6] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: Shape completion and animation of people. *ACM Trans. Graph.*, 24(3):408–416, July 2005. 2

[7] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. V. Guttag. Synthesizing images of humans in unseen poses. In *Computer Vision and Pattern Regognition (CVPR)*, 2018. 3, 4

[8] 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 B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 561–578, Cham, 2016. Springer International Publishing. 2

[9] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 7

[10] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros. Everybody dance now. *arXiv preprint arXiv:1808.07371*, 2018. 1, 2, 3, 6, 7

[11] K.-J. Choi and H.-S. Ko. On-line motion retargeting. 11:223–235, 12 2000. 2

[12] A. Criminisi, P. Perez, and K. Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on Image Processing*, 13(9):1200–1212, September 2004. MSR-TR-2003-83. 6

[13] A. Dosovitskiy, J. T. Springenberg, and T. Brox. Learning to generate chairs with convolutional neural networks. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1538–1546, June 2015. 3

[14] 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. 1, 2, 3, 6

[15] M. Gleicher. Retargetting motion to new characters. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '98, pages 33–42, New York, NY, USA, 1998. ACM. 2

[16] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017. 4, 5, 6

[17] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Regognition (CVPR)*, 2018. 1, 2, 4

[18] S. B. Kang and H.-Y. Shum. A review of image-based rendering techniques. Institute of Electrical and Electronics Engineers, Inc., June 2000. 3

[19] C. Lassner, G. Pons-Moll, and P. V. Gehler. A generative model for people in clothing. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 2, 3

[20] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017. 2

[21] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, Oct. 2015. 2, 3

[22] 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*, pages 405–415, 2017. 2, 3

[23] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. pages 506–516, 10 2017. 4

[24] X. B. Peng, A. Kanazawa, J. Malik, P. Abbeel, and S. Levine. Sfv: Reinforcement learning of physical skills from videos. *ACM Trans. Graph.*, 37(6), Nov. 2018. 1, 2

[25] P. Pérez, M. Gangnet, and A. Blake. Poisson image editing. *ACM Trans. Graph.*, 22(3):313–318, July 2003. 6

[26] S. Saito, , Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *arXiv preprint arXiv:1905.05172*, 2019. 4

[27] H.-Y. Shum, S. B. Kang, and S.-C. Chan. Survey of image-based representations and compression techniques. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(11):1020–1037, Nov 2003. 3

[28] L. Sigal, A. Balan, and M. J. Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, pages 1337–1344, USA, 2007. Curran Associates Inc. 2

[29] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017. 7

[30] M. Tatarchenko, A. Dosovitskiy, and T. Brox. Single-view to multi-view: Reconstructing unseen views with a convolutional network. *CoRR*, abs/1511.06702, 2015. 3

[31] R. Villegas, J. Yang, D. Ceylan, and H. Lee. Neural kinematic networks for unsupervised motion retargetting. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[32] C. Wang, H. Huang, X. Han, and J. Wang. Video inpainting by jointly learning temporal structure and spatial details. In *Proceedings of the 33th AAAI Conference on Artificial Intelligence*, 2019. 6

[33] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro. Video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 1, 2, 3, 6

[34] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE TRANSACTIONS ON IMAGE PROCESSING*, 13(4):600–612, 2004. 7

[35] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016. 7

[36] J. Yang, S. Reed, M.-H. Yang, and H. Lee. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 1099–1107, Cambridge, MA, USA, 2015. MIT Press. 3

[37] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative image inpainting with contextual attention. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 6

[38] C. Zhang and T. Chen. A survey on image-based rendering - representation, sampling and compression. 19:1–28, 01 2004. 3

[39] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7

[40] B. Zhao, X. Wu, Z. Cheng, H. Liu, and J. Feng. Multi-view image generation from a single-view. *CoRR*, abs/1704.04886, 2017. 2