Transferring Human Motion and Appearance in Monocular Videos

Thiago L. Gomes*, Renato Martins*†, Erickson R. Nascimento*
*Department of Computer Science, Universidade Federal de Minas Gerais (UFMG), Brazil
†ImViA, Université Bourgogne Franche-Comté, France
E-mails: {thiagoluange,renato.martins,erickson}@dcc.ufmg.br

***Abstract*—This thesis[1] investigates the problem of transferring human motion and appearance from video to video preserving motion features, body shape, and visual quality. 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. Possible application domains are on graphics animations and entertainment media that rely on synthetic characters and virtual environments to create visual content. We introduce two novel methods for transferring appearance and retargeting human motion from monocular videos, and by consequence, increase the creative possibilities of visual content. Differently from recent appearance transferring methods, our approaches take into account 3D shape, appearance, and motion constraints. Specifically, our first method is based on a hybrid image-based rendering technique that exhibits competitive visual retargeting quality compared to state-of-the-art neural rendering approaches, even without computationally intensive training. Then, inspired by the advantages of the first method, we designed an end-to-end learning-based transferring strategy. Taking advantages of both differentiable rendering and the 3D parametric model, our second data-driven method produces a fully 3D controllable human model, *i.e.*, the user can control the human pose and rendering parameters. Experiments on different videos show that our methods preserve specific features of the motion that must be maintained (*e.g.*, feet touching the floor, hands touching a particular object) while holding the best values for appearance in terms of Structural Similarity (SSIM), Learned Perceptual Image Patch Similarity (LPIPS), Mean Squared Error (MSE), and Fréchet Video Distance (FVD). We also provide to the community a new dataset composed of several annotated videos with motion constraints for retargeting applications and paired motion sequences from different characters to evaluate transferring approaches.**

I. Introduction

Virtual human characters and environments are fundamental components in graphics' animations and in the creation of visual content. Nevertheless, creating these components requires a large amount of manual work wherein artists apply low-level instructions such as drawing the skeletons, manipulating polygons, edges, and vertices for defining realistic human appearance and motions. Due to its importance and wide range of applications, several methods have been proposed on human body reenactment of virtual human characters [1]–[6]. The ultimate goal of these methods is to create a video where the body of a target person is reenacted according to the motion extracted from the monocular video. The motion is estimated considering the set of poses of a source person. Despite the impressive results for several input conditions, there are instances where most of these methods perform poorly. For instance, the works of Chan *et al.* [4] and Wang *et al.* [7], only perform good reenacting of the appearance/style from one actor to another if a strict setup has complied, *e.g.*, static backgrounds, a large set of motion data of the target person to train, and actors in the same distance from the camera [8].

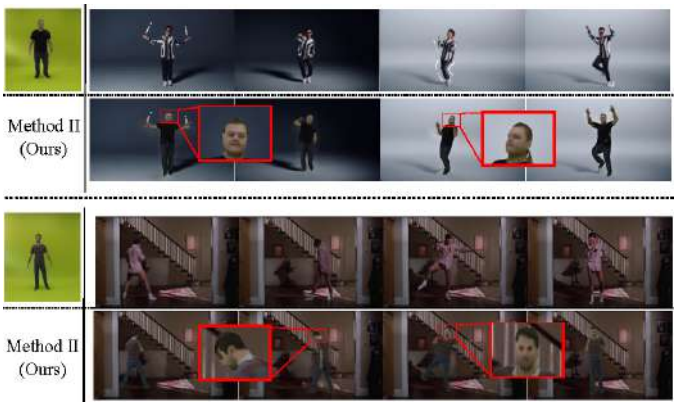

Fig. 1. Human retargeting example. The first line of each scene illustrates the real movement. On the second line is the retargeting using our proposed method. The red squares highlight our face generation quality.

In this thesis, we present two novel video retargeting techniques for human motion and appearance transferring, which incorporate different strategies to extract 3D shape, pose, and appearance to transfer motions between two real human characters using information from monocular videos. To our best knowledge, this work is the first to transfer, not only human texture or motion but both human motion and appearance between videos, *i.e.*, we transfer motion and appearance in a unified way which allows us to tackle subjects with different limb proportions and body shape without losing the desired body proportions. We aim to advance in the task of building a method less sensitive to the camera and poses conditions (a stable method) and overcome the lack of details. Experimental results presented later show that our approaches are both stable and shape-aware. In other words, they do not suffer from quality instability when applied in contexts slightly different from the original ones (a small difference in camera position, uncommon motions, pose translation, etc.) and they can handle different morphologies in

[1]This work relates to a Ph.D. thesis.

the retargeting. Moreover, we performed experiments using a newly collected dataset containing several types of motions and actors with different body shapes and heights. Our results show that a technique applying 3D representation of people can still exhibit a competitive quality compared to recent deep learning techniques in generic transferring tests. Our approaches achieved better results compared with end-to-end 2D learning methodologies such as the works of Wang *et al*. [7] and Chan *et al*. [4] in most scenarios for appearance metrics as structural similarity (SSIM), learned perceptual similarity (LPIPS), mean squared error (MSE), and Fréchet Video Distance (FVD).

The main technical contributions of this work are as follows:

- 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;
- A retargeting technique considering physical constraints of the motion in 3D and the image domain; and a new image-based rendering technique that exhibits competitive visual retargeting quality compared to state-of-the-art neural rendering approaches, even without computationally intensive training;
- A novel data-driven formulation for transfer appearance and reenact human actors that produces a fully 3D controllable human model, *i.e*., the user can control the human pose and rendering parameters;
- A dataset comprising several videos with annotated motion restrictions. We demonstrate the effectiveness of our approach quantitatively and qualitatively using sequences from this dataset and publicly available video sequences. The dataset containing several paired motions and virtual actors is also publicly available to the community[2].

II. Related Work

3D human shape and pose estimation. Significant advances have been recently developed to estimate both the human skeleton and 3D body shape from images. Bogo *et al*. [9] 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 it fits an SMPL body model [10] to these joints. Lassner *et al*. [11] used the curated results from SMPLify to train 91 keypoint detectors. Similarly, Kanazawa *et al*. [12] 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. Kolotouros *et al*. [13] combined an optimization method and a deep network to design a method less sensitive to the optimization initialization. Even though their method outperformed the works of Bogo *et al*. [9], Lassner *et al*. [11], and Kanazawa *et al*. [12] regarding 3D joint error and runtime, their bounding box cropping strategy does not allow motion reconstruction from poses, since it frees three-dimensional pose regression from having to localize the person with scale and translation in image space. Moreover, they lack global information and temporal consistency in shape, pose, and human-to-object interactions, which are required in video retargeting with consistent motion transferring.

Mesh reconstruction. Substantial advances have been made in recent years for 3D model estimation from still images. Human mesh reconstruction methods are also increasingly achieving better results as shown in works such as PiFu [14], [15], ARCH [16], or SiCloPe [17]. Despite the impressive results, these methods are limited to estimate static 3D character models, which require additional efforts to create animated virtual characters. In addition to the requirement that 3D models contain a skeleton hierarchy and appropriate skin weights, it is also necessary to fit a garment model into a human model in various poses. Lazova *et al*. [18] automatically predict a full 3D textured avatar, including geometry and 3D segmentation layout for further generation control; however, their method cannot predict fine details and complex texture patterns.

Retargeting motion. Gleicher's [19] seminal work of retargeting motion addressed the problem of transferring motion from one virtual actor to another with different morphologies. Choi and Ko [20] pushed further Gleicher's [19] method by presenting a real-time motion retargeting approach based on inverse rate control.Villegas *et al*. [21] 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*. [22] takes a step towards automatically transferring motion between humans and virtual humanoids. Similarly, Aberman *et al*. [23] proposed a 2D motion retargeting using a high-level latent motion representation. Their 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 variations of the camera's point-of-view.

Synthesizing views. The past five years has witnessed the explosion of neural rendering approaches and GAN. GANs have emerged as promising and effective approaches to deal with the tasks of synthesizing new views against image-based rendering approaches (*e.g*., [24]–[26]). More recently, the synthesis of views is formulated as being a learning problem (*e.g*., [5], [27]–[30]), where a distribution is estimated to sample the new views. In the work of Esser *et al*. [5], 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, learning-based methods are limited to synthesize detailed body parts such as faces.

Recent works such as Aberman *et al*. [31] and Chan *et al*. [4] start applying adversarial training to map 2D poses 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, their synthesized views still have clear gaps in the test time compared with the training time. Wang *et al*. [7] proposed a general video-to-video synthesis framework based on conditional GANs to generate high-resolution and temporally consistent videos of people. Despite the impressive results

[2]https://github.com/verlab/ShapeAwareHumanRetargeting_IJCV_2021

for several inputs, end-to-end learning-based techniques still fail to synthesize the human body's details, such as face and hands. Furthermore, it is worth noting that these techniques focus on transferring style, which leads to undesired distortions when the characters have different morphologies (proportions or body parts' lengths).

In order to overcome these limitations, Wen *et al*. [6] proposed a 3D body mesh recovery module to disentangle the pose and shape; however, their performance significantly decreases when the source image comes from a different domain from their dataset, indicating that they are also affected by poor generalization to camera viewing changes.

Differentiable rendering. As stated in Kato *et al*. [32], Differentiable Rendering (DR) connects 2D and 3D processing methods and allows neural networks to optimize 3D entities while operating on 2D projections. Loper *et al*. [33] introduced an approximate differentiable render that generates derivatives from projected pixels to the 3D parameters. Kato *et al*. [34] approximated the backward gradient of rasterization with a hand-crafted function. Liu *et al*. [35] proposed a formulation of the rendering process as an aggregation function fusing the probabilistic contributions of all mesh triangles with respect to the rendered pixels. Niemeyer *et al*. [36] represented surfaces as 3D occupancy fields and used a numerical method to find the surface intersection for each ray, then they calculate the gradients using implicit differentiation. While these methods achieved high-quality results, they generally require multi-view data collected with calibrated cameras and have high computational cost, notably during the inference/test time. In this work, we propose a carefully designed architecture for human neural appearance transfer, leveraging the new possibilities offered by differentiable rendering techniques to provide a fully controllable 3D human model.

III. Methodology

This section presents two human transferring methods considering the importance of human motion, body shape, and appearance in the retargeting. Unlike most techniques that transfer either appearance [4], [5], [7], [31] or motion independently [21], [22], we present techniques that simultaneously consider body shape, motion retargeting constraints, and human-to-object interactions over time, while retaining visual appearance quality.

A. General Methodology

This subsection details the steps used to design our two new methods to transfer human motion and appearance from video to video. As depicted in the Figure 2, our two methodologies build upon our general methodology composed of four main components:

1) *Human Motion Estimation:* This component estimates the motion of the character performing actions in the source video, where essential aspects of plausible movements, such as a shared coordinate system for all image frames and temporal motion smoothness are ensured;

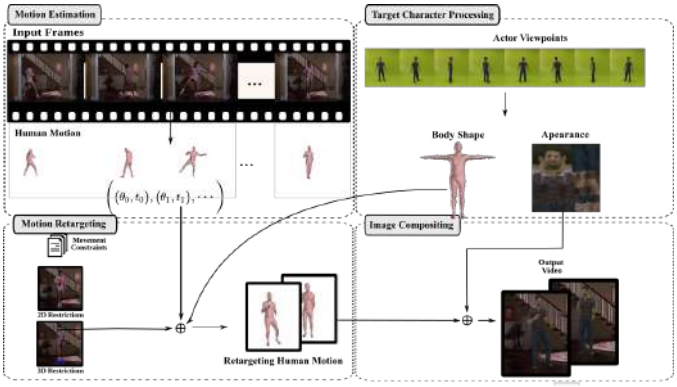

Fig. 2. Overview of our general methodology. Each component is designed to deal with a subproblem of the video-to-video retargeting problem. The four subproblems are: human motion estimation in the source video (Human Motion Estimation); appearance and shape estimation in the target video (Target Character Processing); motion transfer from source character to target character (Motion Retargeting); and target person synthesis into the source video (Compositing).

2) *Target Character Processing:* This component extracts the target character's appearance and body shape in the second video;
3) *Motion Retargeting:* This component adapts the estimated movement to the body shape of the target character while considering temporal motion consistency and the physical human interactions (constraints) with the environment;
4) *Compositing:* This component combines the extracted target character appearance and the adapted movement into the background of the source video.

A central objective of our general methodology is to split the video-to-video retargeting problem into subproblems. Dealing with the subproblems will ensure that our retargeting methods: i) retain the same quality for most poses (Human Motion Estimation); ii) preserve visual quality (Target Character Processing); iii) take into account body shape and the character's interaction with the environment (Motion Retargeting), which allows handling different morphologies in the transferring; and iv) place the target person into a new context (Compositing).

B. Shared Components

In this subsection, we detail the three components shared by our proposed novel video retargeting techniques.

1) Human Motion Estimation: **Human Body and Motion Representation.** To capture the statistics of shape variation and limb-length proportions, we represent the structure of the skeleton together with the 3D shape of the human body using the Skinned Multi-Person Linear (SMPL) model [10] that represents a wide variety of body shapes in natural human poses. The SMPL model ($M(\boldsymbol{\beta}, \boldsymbol{\theta})$) is a skinned vertex-based model, where a mean template mesh of $N = 6890$ vertices is controlled by two sets of parameters, one for body shape ($\boldsymbol{\beta}$), the other for the pose ($\boldsymbol{\theta}$).

Human Pose Model Fitting. Our method builds upon the learning-based SMPL human pose/shape estimation framework of Kolotouros *et al*. [13]. Thus, after cropping the person

using Openpose [37]–[39] and estimating the parameters that represents the 3D reconstruction, we map the reconstruction of Kolotouros *et al*. [13] from the virtual camera coordinates to the original camera by minimizing an objective function that is the sum of two terms: one term that encourages the projections of the joints to remain in same locations into the global reference, and one term that encourages to keep the joints' angles. Together with this process, we force the subject shape to have same mean shape coefficients of the video. Thus, our energy function is given by:

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

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 [13], and λ_1, λ_2 are scaling weights.

Motion Regularization. Since we estimate the character poses frame-by-frame, the resulting motion might present shaking motion with high-frequency artifacts in some short sections of the video. To alleviate these effects, we perform a regularization to seek a new set of joint angles $\widehat{\boldsymbol{\theta}^s}$ that creates a smoother motion. After applying a cubic-spline interpolation [40] over the joints' motion $\mathbf{M}(\boldsymbol{\beta}^s, \boldsymbol{\theta}^s)$, we remove the outlier joints from the interpolated spline. The final motion estimate is obtained by minimizing the cost:

$$\min\Big(||\widehat{\boldsymbol{\theta}^s} - \Theta||_2 + \gamma||\text{FK}(\boldsymbol{\beta}^s, \widehat{\boldsymbol{\theta}^s}) - \mathbf{P}_{sp}||_2\Big), \quad (2)$$

where Θ is the subset of inlier joints, FK is the forward kinematics, $\boldsymbol{\beta}^s$ defines the proportions and dimensions of the human body in the source video, $\mathbf{P}_{sp}$ is the spline interpolated joint positions, and γ is the scaling factor between the original joint angles and the interpolated positions.

2) Motion Retargeting: After estimating the motion from the input video, *i.e*., $\mathbf{M}(\boldsymbol{\beta}^s, \boldsymbol{\theta}^s)$, and 3D model $\boldsymbol{\beta}^t$ of the target human, we can proceed to the motion retargeting step. Our second shared component (*Motion Retargeting*) is essential to guarantee that some physical restrictions are still valid during the target character animation. Similar to Gleicher [19], our first goal is to retain the joint configuration of the target as close as possible to the source joint configurations at instant k, $\boldsymbol{\theta}_k^t \approx \boldsymbol{\theta}_k^s$, *i.e*., to keep $\mathbf{e}_k$ small such as: $\boldsymbol{\theta}_k^t = \boldsymbol{\theta}_k^s + \mathbf{e}_k$. We also aim to keep similar movement style and speed in the retargeted motion. Thus, we propose a one step speed prediction in 3D space defined as $\Delta\mathbf{M}(\boldsymbol{\beta}, \boldsymbol{\theta}_k) = \text{FK}(\boldsymbol{\beta}, \boldsymbol{\theta}_{k+1}) - \text{FK}(\boldsymbol{\beta}, \boldsymbol{\theta}_k)$ to maintain the motion style from the original joints' motion:

$$\mathcal{L}_P(\mathbf{e}) = \sum_{k=i+1}^{i+n} ||\Delta\mathbf{M}(\boldsymbol{\beta}^t, \boldsymbol{\theta}_k^s + \mathbf{e}_k) - \Delta\mathbf{M}(\boldsymbol{\beta}^s, \boldsymbol{\theta}_k^s)||_1, \quad (3)$$

where $\mathbf{e} = [\mathbf{e}_{i+1}, \ldots, \mathbf{e}_{i+n}]^T$, and n is the number of frames considered in the retargeting.

Rather than considering a loss for the total number of frames, we use 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

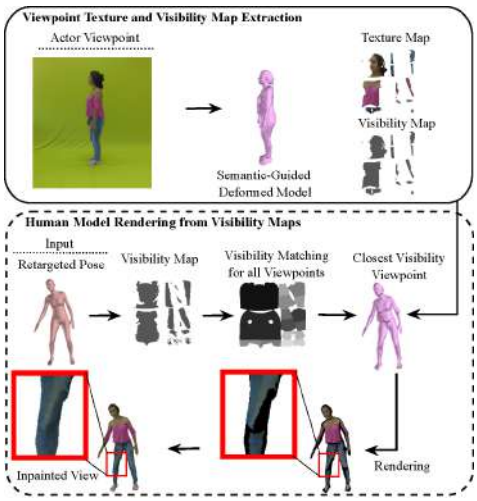

Fig. 3. Rendering of the visibility maps and texture images. *Top:* We project each target actor viewpoint in a common UV texture space using the estimated geometry and create a binary map of visibility body parts. *Bottom:* Given the goal pose (retargeted pose), we estimate its visibility body parts map, and then select the better matching visibility body parts created from the viewpoints from the target actor.

estimated source motion. The retargeting considering a local neighboring window of frames also results in a more efficient optimization.

a) 2D/3D human-to-object interactions.: Going one step further than classic retargeting constraints defined in Gleicher [19] and Choi *et al*. [20], where end-effectors must be at solely a desired 3D position at a given moment, we propose an extended hybrid constraint in the image domain by defining the motion retargeting constraints losses in respect to end-effectors' (hands, feet) 3D poses $\mathbf{P}_{R3D}$ and 2D poses $\mathbf{P}_{R2D}$ as:

$$\mathcal{L}_{R3D}(\mathbf{e}_k) = ||\text{FK}(\boldsymbol{\beta}^t, \boldsymbol{\theta}_k^s + \mathbf{e}_k) - \mathbf{P}_{R3D}||_1, \quad (4)$$

$$\mathcal{L}_{R2D}(\mathbf{e}_k) = ||\Pi(\text{FK}(\boldsymbol{\beta}^t, \boldsymbol{\theta}_k^s + \mathbf{e}_k), \mathbf{K}) - \mathbf{P}_{R2D}||_1. \quad (5)$$

where the $\Pi(., \mathbf{K})$ operator performs the projection taking a 3D point (x, y, z) and projecting it into the image plane given the camera parameters $\mathbf{K}$.

b) Space-time loss optimization: The final motion retargeting loss $\mathcal{L}$ combines the source motion appearance with the different shape and constraints of the target character from Equations 3, 4, and 5:

$$\mathcal{L} = ||\mathbf{W}_1\mathbf{e}||_2 + \lambda_1\mathcal{L}_P(\mathbf{e}) + \lambda_2\mathcal{L}_{R3D}(\mathbf{e}) + \lambda_3\mathcal{L}_{R2D}(\mathbf{e}), \quad (6)$$

where the joint parameters to be optimized are $\mathbf{e} = [\mathbf{e}_{i+1}, \ldots, \mathbf{e}_{i+n}]^T$, n is the number of frames considered in the retargeting window, λ_1, λ_2, and λ_3 are the contributions for the different error terms, and $\mathbf{W}_1$ is a positive diagonal matrix of weights for the motion appearance for each body joint. This weight matrix is set to penalize more errors in joints that are closer to the root joint.

3) Compositing: The third shared component is to compose the final image with the transferred person and the source background. 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*. [41] to ensure temporal smoothness to the final inpainting. We compute the final pixel color value as the median value between the neighboring frames. Finally, the background and the target character are combined in the retargeted frame.

C. Method I: Image-Based Rendering

2D human neural rendering approaches [4], [7], [31], [42] appeared as effective approaches for human appearance synthesis. However, these methods still suffer in creating fine texture details, notably in some body parts as the face and hands. Besides, it is well known that these methods suffer from quality instability when applied in contexts slightly different from the original ones, *i.e*., a small difference in camera position, uncommon motions, pose translation, *etc*. These limitations motivate the proposal of our *Image-Based Rendering* method, which is designed to leverage visibility map information and semantic body parts to refine the initial target mesh model while keeping finer texture details in the transferring.

a) Target Character Processing.: In order to create a more stable method and overcome the lack of details, we design a new semantic-guided image-based rendering approach that copies local patterns from input images to the correct position in the generated images. Our idea stems from using semantic information of the body (*e.g*., face, arms, torso locations, *etc.*) in the geometric rendering to encode patch positions and image-based rendering to copy pixels from the target images, and therefore maintaining texture details. We assert which mesh points are visible by exploring the visibility maps, as illustrated in Figure 3. Each visibility map indicates which parts of the body model are visible per frame. Then we select the closest viewpoint to the desired new viewpoint, for each part of the body model from the visibility maps.

D. Method II: 3D Differentiable Human Rendering

Image-based rendering techniques like our previous technique are effective solutions to create 3D texture-mapped models of people, capable of synthesizing images from any arbitrary viewpoint without using a large number of images. On the other hand, image-based rendering methods cannot improve the visual quality of the synthesized images by using more data when available. Furthermore, the deformation process proposed in our previous method is not fast enough to be used in real-time applications. Thus, we offer a strategy to take advantage of all available data and, in addition, reduce inference time at the cost of increasing preprocessing time (training time).

a) Target Character Processing: In order to generate a deformable 3D texture-mapped human model, our end-to-end architecture has three main components to be trained during the rendering. The first component models local deformations on the human 3D body shape extracted from the images using a three-stage GCN. In the second component, a CNN is trained to estimate the human appearance map. Similar to the GCN, the CNN is also trained in a self-supervised regime using the gradient signals from the differentiable renderer. Finally, the third component comprises an adversarial regularization in the human appearance texture domain to ensure the reconstruction of photo-realistic images of people. In the inference/test time, we can feed our architecture with generic meshes parametrized by the SMPL model, and then we can create a refined mesh and a detailed texture map to represent the person's shape and appearance properly. Figure 4 outlines these components during the inference phase.

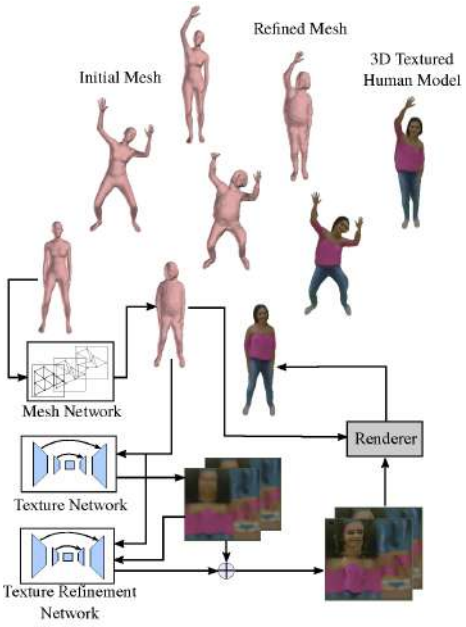

Fig. 4. Overview of our data-driven formulation for transfer appearance and reenact human actors. Our method receives a set of frames of a person, extracts her/his mesh (left side) and outputs a fully 3D controllable human model (right side).

IV. Human Retargeting Dataset

To evaluate the retargeting and appearance transfer with different actor motions, consistent reconstructed 3D motions, and with human-to-object interactions, we created a new dataset with *paired motion* sequences from different characters and *annotated motion retargeting constraints*. For each video sequence, we provide a refined 3D actor reconstructed motion and the actor body shape estimated [43]. The refined reconstructed 3D motions and 2D-3D annotation of interactions were collected by manual annotation. Figure 5 shows some examples of frames from our dataset.

V. Experiments and Results

We compare our methods against four recent methods including V-Unet [5], Vid2Vid [7], EBDN [4] and the Impersonator [6]. We adopted complementary metrics to evaluate the quality of the approaches to asset different aspects of the generated images such as structure coherence, luminance, contrast, perceptual similarity [44], temporal, and spatial coherence. The metrics used to perform quantitative analysis are SSIM [45], LPIPS [44], Mean Square Error (MSE), and

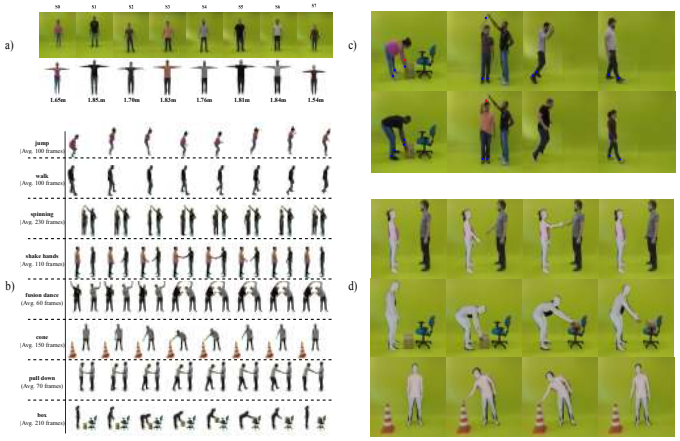

Fig. 5. **Human retargeting dataset.** *a)* The subjects participating in our dataset, their respective height and estimated SMPL body models. *b)* Overview of all motions available in our proposed dataset. *c)* Paired motions (upper and lower rows) with annotated motion constraints (3D constraints in blue and 2D constraints in red). *d)* The reconstructed 3D motions.

Fig. 6. Qualitative comparison. Transferring results considering the cases where the person is not standing parallel to the image plane or has the arms in front of the face. In each sequence: the first row shows the worst generated frame for each method and the second row presents the best generated frame for each method.

TABLE I
Comparison with state of the art. AVERAGE SSIM, LPIPS, MSE, AND FVD (BEST IN BOLD, SECOND-BEST IN ITALIC).

Metric	Method					
	V-Unet	vid2vid	EBDN	iPER	Method I (Ours)	Method II (Ours)
SSIM↑	0.849	0.862	0.861	0.859	*0.864*	**0.868**
LPIPS↓	0.167	0.138	0.153	0.167	*0.135*	**0.134**
MSE↓	368.44	303.32	334.63	350.23	*274.11*	**259.79**
FVD↓	1,639.09	*708.72*	732.60	1,243.42	**651.10**	712.11

Fréchet Video Distance (FVD) [46]. We executed all the methods in the motion sequences and transferred them to the same background. This protocol allows us to generate comparisons with the ground truth and compute the metrics for all the generated images with their respective real peers.

A. Quantitative Comparison with State of The Art

We performed the human retargeting for actors with different body shapes, gender, clothing styles, and sizes for all considered video sequences. The video sequences used in the actor animation contained motions with different levels of difficulty, which aims to test the generalization capabilities of the methods in unseen data. Table I shows the performance for each method considering all paired videos in the dataset. We can see that our Method II achieves superior performance as compared to the methods in most metrics.

B. Qualitative Visual Analysis

The visual inspection of synthesized actors also concur with the quantitative analysis. In Figure 6, we provide the worst/best frames for each movement using four actors in the dataset. Our Method I and our Method II are the only models capable of keeping the body scale of the authors along all scenes, while the other methods failed, in particular in the movements *shake hands* and *walk*. Besides generating coherent poses, our second method also generated more realistic textures in comparison to the other methods. Comparing the results of the movements *jump* and *spinning*, one can visualize some details as the shadow of the shirt sleeve of the actor and the shirt collar, respectively. The Figure 1 illustrates a task of retargeting in two different scenarios. These results demonstrate the capability of generating detailed face and body texture, producing a good fit of the actors in the different scenes.

VI. CONCLUSION

This thesis proposes a general methodology of transferring human motion and appearance from video to video preserving motion features, body shape, and visual quality. We designed two novel methods using our proposed general methodology and we demonstrated that this methodology is adequate to be used as a design guide in the creation of new methods to transfer human motion and appearance from video to video preserving motion features, body shape, and visual quality. From a theorical standpoint, our work exploits motion constraints, body shape, and a 3D representation of people to synthesizing more plausible videos and allows us to tackle subjects with different limb proportions and body shape.

Acknowledgments: We would like to thank the PPGCC-UFMG, CAPES, FAPEMIg, and CNPq for funding different parts of this work.

VII. PUBLICATIONS AND AWARDS

The results of this dissertation were published in the *International Journal of Computer Vision* (IJCV'21) [47], and in two international conferences on applications of computer vision (WACV'20 [48] and WACV'22 [49]). The student also contributed as co-author to one related publication on human motion generation in the international journal *Computers & Graphics* [50]. We also would like to highlight that this PhD work has been selected as the best PhD work of the Department of Computer Science of UFMG and was appointed to the CAPES and UFMG thesis award in 2021.

REFERENCES

[1] C. Lassner, G. Pons-Moll, and P. V. Gehler, "A generative model for people in clothing," in *CVPR*, 2017.

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

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

[4] C. Chan, S. Ginosar, T. Zhou, and A. Efros, "Everybody dance now," in *ICCV*, 2019.

[5] P. Esser, E. Sutter, and B. Ommer, "A variational u-net for conditional appearance and shape generation," in *CVPR*, 2018.

[6] W. Liu, Z. Piao, M. Jie, W. Luo, L. Ma, and S. Gao, "Liquid warping GAN: A unified framework for human motion imitation, appearance transfer and novel view synthesis," in *ICCV*, 2019.

[7] 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.

[8] A. Tewari, O. Fried, J. Thies, V. Sitzmann, S. Lombardi, K. Sunkavalli, R. Martin-Brualla, T. Simon, J. Saragih, M. Niebner, R. Pandey, S. Fanello, G. Wetzstein, J.-Y. Zhu, C. Theobalt, M. Agrawala, E. Shechtman, D. B. Goldman, and M. Zollhofer, "State of the art on neural rendering," *Computer Graphics Forum*, vol. 39, no. 2, pp. 701–727, 2020.

[9] 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 *ECCV*, 2016.

[10] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," *ACM Trans. Graph.*, 2015.

[11] 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 *CVPR*, 2017.

[12] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *CVPR*, 2018.

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

[14] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, "PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization," in *IEEE/CVF International Conference on Computer Vision*, 2019.

[15] S. Saito, T. Simon, J. Saragih, and H. Joo, "PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[16] Z. Huang, Y. Xu, C. Lassner, H. Li, and T. Tung, "ARCH: animatable reconstruction of clothed humans," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[17] R. Natsume, S. Saito, Z. Huang, W. Chen, C. Ma, H. Li, and S. Morishima, "Siclope: Silhouette-based clothed people," in *CVPR*, 2019.

[18] V. Lazova, E. Insafutdinov, and G. Pons-Moll, "360-degree textures of people in clothing from a single image," in *2019 International Conference on 3D Vision, 3DV 2019, Québec City, QC, Canada, September 16-19, 2019*. IEEE, 2019, pp. 643–653.

[19] M. Gleicher, "Retargetting motion to new characters," in *SIGGRAPH*, 1998.

[20] K.-J. Choi and H.-S. Ko, "On-line motion retargeting," *Journal of Visualization and Computer Animation*, 2000.

[21] R. Villegas, J. Yang, D. Ceylan, and H. Lee, "Neural kinematic networks for unsupervised motion retargetting," in *CVPR*, June 2018.

[22] X. B. Peng, A. Kanazawa, J. Malik, P. Abbeel, and S. Levine, "Sfv: Reinforcement learning of physical skills from videos," *ACM Trans. Graph.*, vol. 37, no. 6, Nov. 2018.

[23] K. Aberman, R. Wu, D. Lischinski, B. Chen, and D. Cohen-Or, "Learning character-agnostic motion for motion retargeting in 2d," *ACM TOG*, 2019.

[24] S. B. Kang and H.-Y. Shum, "A review of image-based rendering techniques," 2000.

[25] C. Zhang and T. Chen, "A survey on image-based rendering-representation, sampling and compression," *Signal Processing: Image Communication*, 2004.

[26] H.-Y. Shum, S. B. Kang, and S.-C. Chan, "Survey of image-based representations and compression techniques," *TCSVT*, 2003.

[27] M. Tatarchenko, A. Dosovitskiy, and T. Brox, "Single-view to multi-view: Reconstructing unseen views with a convolutional network," *CoRR*, vol. abs/1511.06702, 2015. [Online]. Available: http://arxiv.org/abs/1511.06702

[28] A. Dosovitskiy, J. T. Springenberg, and T. Brox, "Learning to generate chairs with convolutional neural networks," in *2015 IEEE Conference on CVPR*, June 2015, pp. 1538–1546.

[29] 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*, ser. NIPS'15. Cambridge, MA, USA: MIT Press, 2015, pp. 1099–1107. [Online]. Available: http://dl.acm.org/citation.cfm?id=2969239.2969362

[30] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. V. Guttag, "Synthesizing images of humans in unseen poses," in *CVPR*, 2018.

[31] K. Aberman, M. Shi, J. Liao, D. Lischinski, B. Chen, and D. Cohen-Or, "Deep video-based performance cloning," *CoRR*, 2018.

[32] H. Kato, D. Beker, M. Morariu, T. Ando, T. Matsuoka, W. Kehl, and A. Gaidon, "Differentiable rendering: A survey," *arXiv preprint*, 2020.

[33] M. M. Loper and M. J. Black, "Opendr: An approximate differentiable renderer," in *European Conference on Computer Vision*, 2014.

[34] H. Kato, Y. Ushiku, and T. Harada, "Neural 3d mesh renderer," in *CVPR*, 2018.

[35] S. Liu, T. Li, W. Chen, and H. Li, "Soft rasterizer: A differentiable renderer for image-based 3d reasoning," *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019.

[36] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger, "Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision," in *Proceedings of the IEEE/CVF Conference on CVPR*, June 2020.

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

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

[39] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *CVPR*, 2016.

[40] C. De Boor, C. De Boor, E.-U. Mathématicien, C. De Boor, and C. De Boor, *A practical guide to splines*, 1978, vol. 27.

[41] A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE TIP*, 2004.

[42] Y.-T. Sun, Q.-C. Fu, Y.-R. Jiang, Z. Liu, Y.-K. Lai, H. Fu, and L. Gao, "Human motion transfer with 3d constraints and detail enhancement," 2020.

[43] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, "Video based reconstruction of 3d people models," in *CVPR*, 2018.

[44] 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.

[45] 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*, vol. 13, no. 4, pp. 600–612, 2004.

[46] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards accurate generative models of video: A new metric & challenges," 2019.

[47] T. L. Gomes, R. Martins, J. Ferreira, R. Azevedo, G. Torres, and E. R. Nascimento, "A shape-aware retargeting approach to transfer human motion and appearance in monocular videos," *International Journal of Computer Vision*, Apr 2021. [Online]. Available: https://doi.org/10.1007/s11263-021-01471-x

[48] T. L. Gomes, R. Martins, J. P. Ferreira, and E. R. Nascimento, "Do as i do: Transferring human motion and appearance between monocular videos with spatial and temporal constraints," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Aspen, USA, 2020.

[49] T. L. Gomes, T. M. Coutinho, R. Azevedo, R. Martins, and E. R. Nascimento, "Creating and reenacting controllable 3d humans with differentiable rendering," in *WACV*. IEEE, 2022, pp. 717–726.

[50] J. P. Ferreira, T. M. Coutinho, T. L. Gomes, J. F. Neto, R. Azevedo, R. Martins, and E. R. Nascimento, "Learning to dance: A graph convolutional adversarial network to generate realistic dance motions from audio," *Computers & Graphics*, vol. 94, pp. 11 – 21, 2021.