Human Action Transfer Based on 3D Model Reconstruction

Shanyan Guan,[1*] **Shuo Wen,**[1*] **Dexin Yang,**[1*] **Bingbing Ni,**[1,2,3†]
Wendong Zhang,[1] **Jun Tang,**[1,2] **Xiaokang Yang**[1,2,3]
[1]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[2]SJTU-UCLA Joint Research Center on Machine Perception and Inference
[3]Shanghai Institute for Advanced Communication and Data Science
{shyanguan, wenshuo, HaDean1998, nibingbing, diergent, tangjun1994, xkyang}@sjtu.edu.cn

Abstract

We present a practical and effective method for *human action transfer*. Given a sequence of source action and limited target information, we aim to transfer motion from source to target. Although recent works based on GAN or VAE achieved impressive results for action transfer in 2D, there still exists a lot of problems which cannot be avoided, such as distorted and discontinuous human body shape, blurry cloth texture and so on. In this paper, we try to solve these problems in a novel 3D viewpoint. On the one hand, we design a skeleton-to-3D-mesh generator to generate the 3D model, which achieves huge improvement on appearance reconstruction. Furthermore, we add a temporal connection to improve the smoothness of the model. On the other hand, instead of directly utilizing the image in RGB space, we transform the target appearance information into UV space for further pose transformation. Specially, unlike conventional graphics render method directly projects visible pixels to UV space, our transformation is according to pixel's semantic information. We perform experiments on Human3.6M and HumanEva-I to evaluate the performance of pose generator. Both qualitative and quantitative results show that our method outperforms methods based on generation method in 2D. Additionally, we compare our render method with graphic methods on Human3.6M and People-snapshot. The comparison results show that our render method is more robust and effective.

Introduction

Recently, human action comprehension (such as action recognition and detection) has attracted great interest and attention from academic society and there are plentiful remarkable achievements (Sigurdsson, Russakovsky, and Gupta 2017; Fan et al. 2018; Yang, He, and Porikli 2018; Zhou et al. 2018). However, researches on human action transfer are still in progress. In terms of human action transfer, given the appearance of the target person and a video (motion sequence) of the source person, our task is to impose the motion of the source person on the target person.

With the development of deep generative model, there have been several works that can perform human action transfer in both image level (Esser, Sutter, and Ommer 2018) and video level (Chan et al. 2018). All of these works are built on the image to image translation architecture, which tries to solve this problem only in 2D space.

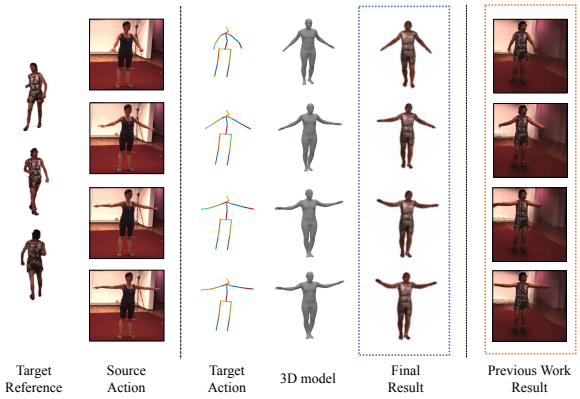

Figure 1: Given the appearance of the target person and a video of the source person, our model impose the motion of the source person on the target person.

However, this task is still very challenging due to two major reasons: (1) Discontinuity of body shape between adjacent frames and between distant frames often occurs. (2) Detailed texture is difficult to be preserved when using image generation based methods such as GAN or VAE. For example, the weight of the target's arm may change in different frames; the pattern printed on clothes will be blurred. Both of these difficulties are caused by that image generation based methods have an upper bound on the performance of generation quality due to its intrinsic large searching space on high dimensions. Large number of data is needed for these methods to learn appearance and pose information and to generate high-quality image.

In this paper, we overcome the aforementioned problems by proposing a novel framework for human action transfer based on 3D human body model generation, as shown in Figure 1. Our approach differs from previous works by adopting the 3D human model construction method based on SMPL model (Loper et al. 2015), which narrows the searching space to 82 shape and pose parameters. With this method, we can easily map the appearance images and pose

[*]Equal contribution
[†]Corresponding author: Bingbing Ni

skeletons to a realistic 3D human model, while maintaining the completeness and consistency of human body shape.

To preserve the detailed texture of the target person, we propose a new method for texture extraction which only needs 3 images of the target person on different views. Moreover, different from methods like (Chan et al. 2018) which need a distinct model trained for each target person with skeleton-image paired dataset, our method to extract and render texture is easily generalized to different target person without additional training.

Furthermore, as for the texture extraction and rendering step, there have been many works using graphics methods (Alldieck et al. 2018), which often require the target person to be at the fixed pose (such as A-pose or T-pose) to avoid occlusion problems. Our method, however, can work with the target person at various poses with the help of semantic map provided by DensePose (Güler et al. 2018), which enables us to accurately determine which part of the human that each pixel belongs to

We apply our method on Human3.6M dataset (Ionescu et al. 2014; Ionescu 2011) and HumanEva dataset (Sigal, Balan, and Black 2009), then evaluate it with the structural similarity (SSIM) (Wang et al. 2004) and Peak Signal-to-Noise Ratio (PSNR). The experimental results indicate competitive performance of our method compared to image generation based methods.

As a summary, the contribution of this paper is as follows:

- We propose a novel framework for human action transfer based on 3D human body model generation, which maintains human shape completeness and consistency.
- We propose a new method for texture extraction to preserve detailed texture with only 3 appearance images of different views.
- Our framework can be easily generalized to perform action transfer on various target person without additional training.

Related Work

Human action video generation based on GAN or VAE has two common problems: discontinuous body shape and blurring texture. Although the retargeting effect of (Chan et al. 2018) works well, the structure of the hands and feet will still be distorted or disappeared in the demo they released a few days ago. To preserve human body structure, we adopt parametric 3D model generation which has strong structure prior. To overcome occlusion problem, we generate human texture in UV space using three images of different views, and render them using barycentric transform. Here we review action transfer or retargeting method and 3D human shape generation and clothing for monocular images.

Motion Retargeting. This task is first introduced in 3D animated characters (Gleicher 1998; Choi 2000; Tak and Ko 2005), where they retargeted motion between two 3D characters by applying forward kinematics (FK) with inverse kinematics (IK) techniques and space constraints. Extending these techniques, motion retargeting can be performed on more varied user-created morphologies (Hecker et al. 2008; Bellini, Kleiman, and Cohen-Or). Most works mentioned above regard motion retargeting task as an optimization problem and solve them with hand-designed kinematic constraints for particular motions. Recently, (Villegas et al. 2018) used deep learning methods and achieved unsupervised motion (skeleton) retargeting by applying cycle consistency objective.

3D Human Shape Generation and Rendering. Most of 3D human model generation works are based on SMPL, which is a parametric model decided by shape (β) and pose (θ). (Bogo et al. 2016) reconstruct 3D human model with 2D joints constraints. However, reconstructing a 3D model with only a single frame of 2D skeleton information often results in opposite results. Thus, (Huang 2017) use multi-view 2D joints and adds discrete cosine transform to deal with it. Most of 3D reconstruction works (Lassner et al. 2017; Kanazawa et al. 2018; Zanfir et al. 2018; Pavlakos et al. 2018) are aimed to estimate 3D shape for image. (Alldieck et al. 2018) propose a new method oriented to video sequence. More importantly, they propose a render method based on graphics. Although many works have made progress in reconstructing 3D models from images or videos, but as far as we know, there is no work dedicated to transforming skeletons to 3D human models.

Method

Different from (Chan et al. 2018), we focus on using 3D human model (SMPL) to solve the problem of body shape distortion or even disappearance. SMPL is a triangulated naked mesh model with $N = 6890$ vertices. It is decided by two sets of parameters, which are *shape* $\beta \in R^{10}$ (representing individual's weight, height, etc.) and *pose* $\theta \in R^{72}$ (representing relative rotation of $K = 17$ 3D joints in kinematics tree). The deformations generated by β and θ are linearly superimposed on a statistical mean model T_μ, showed in Eq 1. As shown in Eq 2, based on 3D joints J, SMPL can generate various of models with different θ and β through a linear blend-skinning function.

$$T(\beta, \theta) = T_\mu + B_s(\beta) + B_p(\theta), \tag{1}$$

$$M(\beta, \theta) = W(T(\beta, \theta), J, \theta, \mathbf{W}), \tag{2}$$

where $T(\beta, \theta)$ is a deformed template; W is a linear blend-skinning function; $\mathbf{W}$ is the function weight. Meanwhile, 3D joints J can also be obtained by linear regression from the final vertices $M(\beta, \theta)$,

$$J = R(\beta, \theta), \tag{3}$$

where $J \in R^{23\times 3}$, and R is a linear regression function.

Figure 2 shows the overview of our pipline. Our pipline consists of three steps: (1) *skeleton transfer*, (2) *3D model generation* and (3) *texture rendering*.

At the first step, we utilize DMHS (Popa, Zanfir, and Sminchisescu 2017) to estimate human's 2D and 3D joints. Considering that the joint angle and bones length of the skeleton are the main factors affecting the action, we adjust the joint angle and bones length to achieve action transfer.

At the second step, due to the fact that SMPL is linearly affected by θ and β, we assume that the same person should

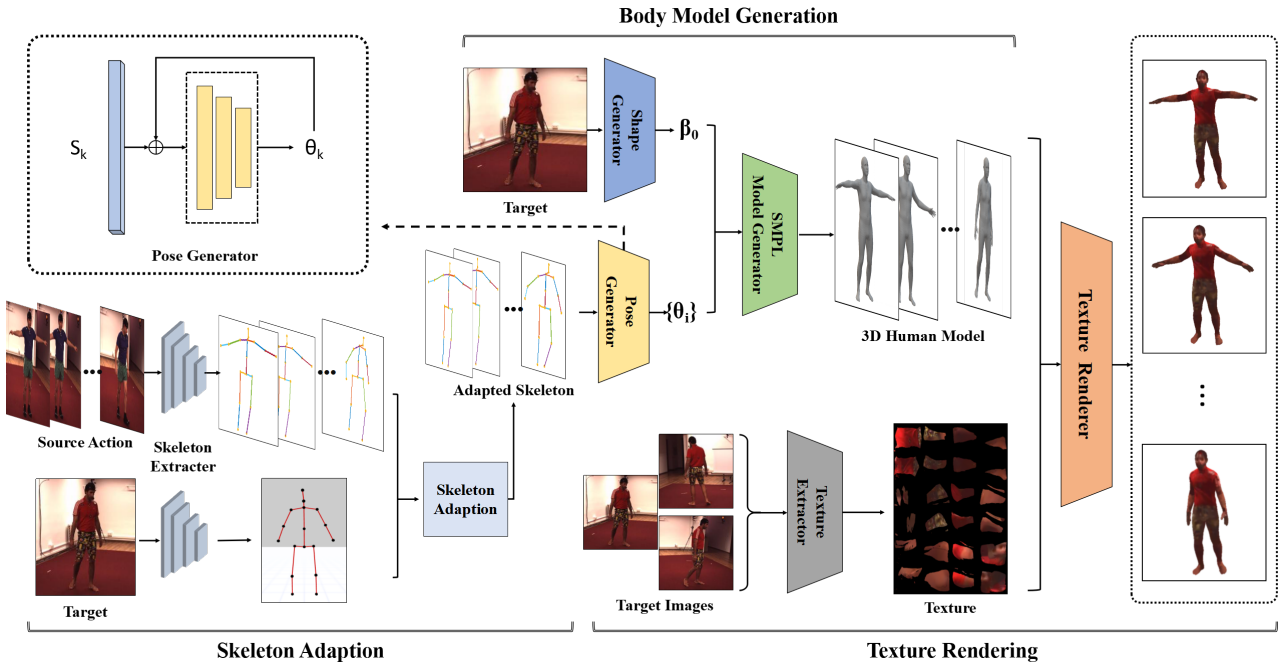

Figure 2: The overall framework for generating the 3D model which have the appearance of the target person and the motion of the source person. First we do the skeleton transfer by combining the skeleton bones length of the target person and the pose information of the source person. Then we generate the 3D human model (SMPL) based on the transformed skeleton and the shape of the target person. After that, we obtain the texture from the three pictures of the target person. Finally, using the method of barycentric transformation, we render the texture on the 3D human model.

have the same β. Firstly, we randomly choose an image from three reference images, and load a pretrained model provided by HMR (Kanazawa et al. 2018) to estimate initial shape β_0. Then we design another network to encode transferred skeleton sequence into corresponding pose parameters. Particularly, to ensure temporal consistence, we concatenate current skeleton S_t and the former timestep pose θ_{t-1} into network.

At the third step, we aim to render human's cloth to 3D mesh model with different poses. According to information theory. The more pictures are provided, the less uncertain the information is. In our experiment, we found that the three images is enough.

Action Transfer

Since our goal is to combine the target's appearance with the source action, we separate the information of skeleton bone lengths from pose of skeletons by using kinematic tree, in which the pose of a skeleton is expressed recursively, starting from the root joint and ending in the leaf joints. Specifically, each skeleton can be expressed by pose D and skeleton bone lengths L. Then the position of the nth joint p_n can be expressed as

$$p_n = p_{parent(n)} + D_n L_n, \tag{4}$$

where $p_{parent(n)}$ means the position of parent node of the nth joint; D_n is the direction of the nth joint with respect to its parent; L_n is the length between the nth joint and its parent, which can be expressed as

$$L_n = ||p_n - p_{parent(n)}||. \tag{5}$$

Therefore, given the root of the kinematic tree and the skeleton information $\{L, D\}$, we can calculate the position of each joints recursively. In this work, we use left foot as the root. And we use 17 joints skeleton as our skeleton model.

For example, as shown in Figure 3, the position of the left hand can be expressed as

$$\begin{aligned} p_{17} &= p_{parent(17)} + L_{17}D_{17} \\ &= p_{parent(16)} + L_{16}D_{16} + L_{17}D_{17} \\ &= \cdots \\ &= p_7 + L_6D_6 + \cdots + L_{16}D_{16} + L_{17}D_{17}. \end{aligned} \tag{6}$$

since the 7th joint is the root (left foot). By this method, we can use the skeleton bone lengths and the kinematic tree to express the position of joints instead of using three-dimensional Cartesian coordinate.

Using this method, we can transform the skeleton to match target's appearance by combining the skeleton bone lengths of the target and the pose information obtained from the motion sequence. First we get the skeleton $\{L_t, D_t\}$ by measuring the appearance of the target person from the given picture. Then, we get the skeleton sequence $\{L_m, D_m^n\}$ from the motion sequence with n frames of the source person. Finally, the result $\{L_t, D_m^n\}$ can be obtained by combining

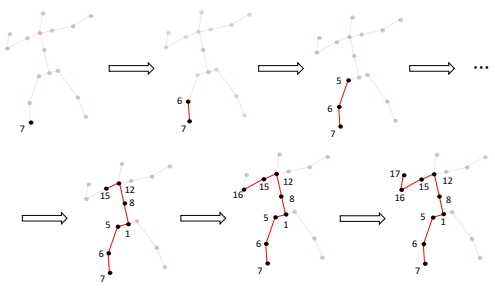

Figure 3: Recursive joints position calculation of the kinematic tree. Starting from the root joint (left foot), this method calculates each joint position recursively.

the target's skeleton bone lengths and the motion sequence's pose.

We also adjust the scale and the overall position of the skeleton sequence for projecting the 3D model, which is based on the skeleton sequence, back into the picture at the end of the whole process. Fortunately, using the skeleton bone lengths of the target, the scale of the skeleton is the same as in the original picture. For the overall position, we view the midpoint of the two feet as the fixed point since under most circumstances people stand on the ground instead of floating in the air.

Body Model Generation

Different from previous 3D people reconstruction methods which use both semantic map and joints of each frame, our model is conditioned on only one reference image I_r and the skeleton sequence $\{S_1, S_2, ...\}$. Our model contain two parts: one is shape generator G_s (generate initial shape parameter β_0) and another one is pose generator G_p (generate a set of pose parameters $\{\theta_1, \theta_2, ...\}$ corresponding to action sequence). Combining β_0 and $\{\theta_1, \theta_2, ...\}$, we can generate a continuous SMPL sequence.

Shape Generator. we adopt the same structure as HMR and load a pretrained model. Inputting reference image I_r, shape generator will output corresponding human shape $\beta \in R^{10}$. The formulation can be represented as follows:

$$\beta = G_s(I_r), \tag{7}$$

where β is the initial shape.

Pose Generator. we adopt three linear layers to get pose. Noticing that our goal is to generate human model with continuous motion rather than individual human model, we concatenate 3D joints $J_t^{3d} \in R^{17\times 3}$ at current timestep with the last timestep pose θ_{t-1} and then input it to G_p. Our network structure is illustrated in Figure 4. We have experimentally verified that this operation achieves better results than directly inputting 3D joints. The process can be express as:

$$\theta_t = G_p([J_t^{3d}, \theta_{t-1}]), \tag{8}$$

for $t \in 1, 2, ..., F$.Here,$[\cdot, \cdot]$ is concatenation operator, and F is the number of action sequence. At each time-step, we

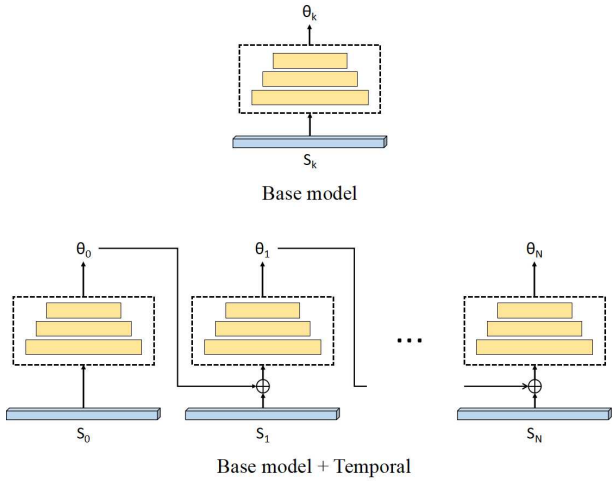

Figure 4: Pose Generator. To ensure the motion continuity, our model takes both the pose information and the temporal information into consideration. *Above:* The basic model, *Below:* Our model.

employ L2 loss on 3D joints and L1 loss on 2D joints to guide pose generator for learning pose better.

$$L_{3Djoints} = \sum_{i=1}^{K} ||J_i^{3D} - J_i^{smpl}||_2^2 \tag{9}$$

$$L_{2Djoints} = \sum_{i=1}^{K} ||J_i^{2D} - \Pi(J_i^{smpl})||_1 \tag{10}$$

$$J^{smpl} = R(\beta_t, \theta_t) \tag{11}$$

where J_i^{smpl} is the ith 3D joints regressed by generated 3D model, and $\Pi(J_i^{smpl})$ is 2D joints projected by an orthographic projection Π, $i = 1...K$ represent joints index.

However, only using these two constraints, our method will produce manlike monster models, as shown in later ablation study section. Referring to SMPLify(Bogo et al. 2016), we adopt pose prior to teach model what distribution pose should obey:

$$L_\theta = min_j(-g_j log(N(\theta; \mu_{\theta,j}, \Sigma_{\theta,j}))) \tag{12}$$

where $N(\mu_j, \Sigma_j)$ is the jth component of Mixture Gaussian distribution; g_j represents its corresponding weight; and μ and Σ are learned from CMU dataset(Akhter and Black 2015).

To sum up, our loss function for 3D generation model is defined as :

$$L_{total} = L_{3Djoints} + \lambda_{2d} L_{2Djoints} + \lambda_\theta L_\theta \tag{13}$$

where λ_{2d} and λ_θ are corresponding loss weight.

Implementation Details. Our implementation is based on Pytorch(Paszke et al. 2017). Shape generator has the same structure as HMR. Pose generator consists of two linear layers with 1024 neurons, followed by an output linear layer. Results detected by DMHS contain 17 joints, which is

not compatible with SMPL joints. For this reason, we choose 14 joints of them to calculate $L_{3Djoints}$ and $L_{2Djoints}$. During training, λ_{2d} is set to 10, and λ_θ is set to 1. We use Adam optimizer(Kingma and Ba 2014) with default setting in Pytorch. For shape generator, we freeze it during training. For pose generator, we use initial learning rate of 1×10^{-4} and decrease it by 0.1 per 3 epoches. Batchsize is set to 64, and training typically took 15 hours on a GPU (TITAN X).

Texture Rendering

With the help of DensePose (Güler et al. 2018), we render the texture on the 3D model. Different from the methods in computer graphics, the pose of the target person will not badly impact on our texture generation since DensePose use semantic map to decide each pixel belongs to which of the body part before corresponding the pixels of 2D image to the 3D model.

Here, we first transfer the RGB image into UV field; then, we find the color information of each point on 3D model from UV field. For the correspondence between 2D image and UV field, DensePose has already performed well on it. However, it can only find the correspondences of the existing areas without predicting the missing areas. Therefore, for obtaining the whole texture, at least 2 images which provide texture information from multi-angle are needed. Besides, we apply image inpainting on it in order to recover the missing area caused by occlusion. After getting a complete texture, we seek the color information of each point on the 3D model from the texture by two steps: first, we find the corresponding position of the 6890 vertices of the 3D model on texture; then, we use barycentric transformation to find the corresponding position of the other points on the 3D model. For barycentric transformation, we first get 13776 triangle areas formed by the 6890 vertices; after that, in each triangle area, we insert a certain number of joints by combining different times of base vectors of the barycentric coordinate; finally, we match the points in each triangle area into UV space by calculating the combination of the corresponding base vectors in UV space.

For example, when we render the texture in triangle abc as shown in Figure 5 , we first match the position of apex a,b,c in UV space, which are shown as $a^{'}$,$b^{'}$,$c^{'}$; then we calculate each point p in the triangle abc by combining m times base vectors $\mathbf{ab}$ and n times $\mathbf{ac}$; finally, we match the position of p to $p^{'}$ in the UV space by combining m times base vectors $\mathbf{a^{'}b^{'}}$ and n times $\mathbf{a^{'}c^{'}}$.

In this work, we insert 65 points in each triangle. The number of points in each triangle can be changed according to the image resolution we need.

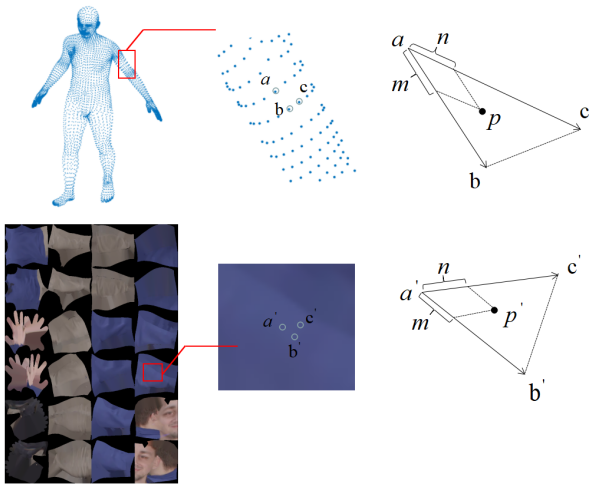

Figure 5: Barycentric transformation. Given the correspondence of the 6890 vertices in UV space and the 13776 triangles formed by the vertices, barycentric transformation corresponds with the points which locate in the 13776 triangles into UV space and then get the color information.

Experiments

Datasets

We evaluate the whole framework on Human3.6M and HumanEva-I dataset, and evaluate the texture rendering step individually in people-snapshot dataset(Alldieck et al. 2018).

Human3.6M Human3.6M dataset is a widely used benchmark for human action and pose related tasks, which contains 3.6 million 3D human poses and corresponding images with 11 professional actors and 17 different scenarios. It provides High-resolution (1000×1000) 50 frames per second video from 4 calibrated cameras and accurate 3D joint positions and joint angles from high-speed motion capture system, which we can use to perform skeleton adaption. In this experiment, we choose videos of five people (3 males and 2 females) as training set and videos of two people (1 male and 1 female) as test set.

HumanEva-I HumanEva-I dataset is a structured comprehensive development dataset for human pose estimation and motion tracking, which contains 7 calibrated video sequences that are synchronized with 3D body poses obtained from a motion capture system with 4 subjects performing 6 common actions. We choose videos of *Walking* and *Box* of all three actors as training set and videos of *Gestures*, *Jog*, *Throw* and *Catch* as test set.

People-snapshot The People-Snapshot dataset (Alldieck et al. 2018) consists of 24 sequences of 11 subjects varying a lot in height and weight. The sequences are captured with HD camera, and the resolution of each frame is 1080×1080. All subjects are required to rotate while holding A-pose.

Evaluation Metrics

To assess the experimental results, we use both structural similarity (SSIM) (Wang et al. 2004) and Peak Signal to Noise Ratio (PSNR) to evaluate the generation quality of transferred video frames. Brenner gradient (Brenner et al. 1976) and Mean Opinion Score (MOS) are also used to assess the quality of generated texture images and the effect of our pose generator structure.

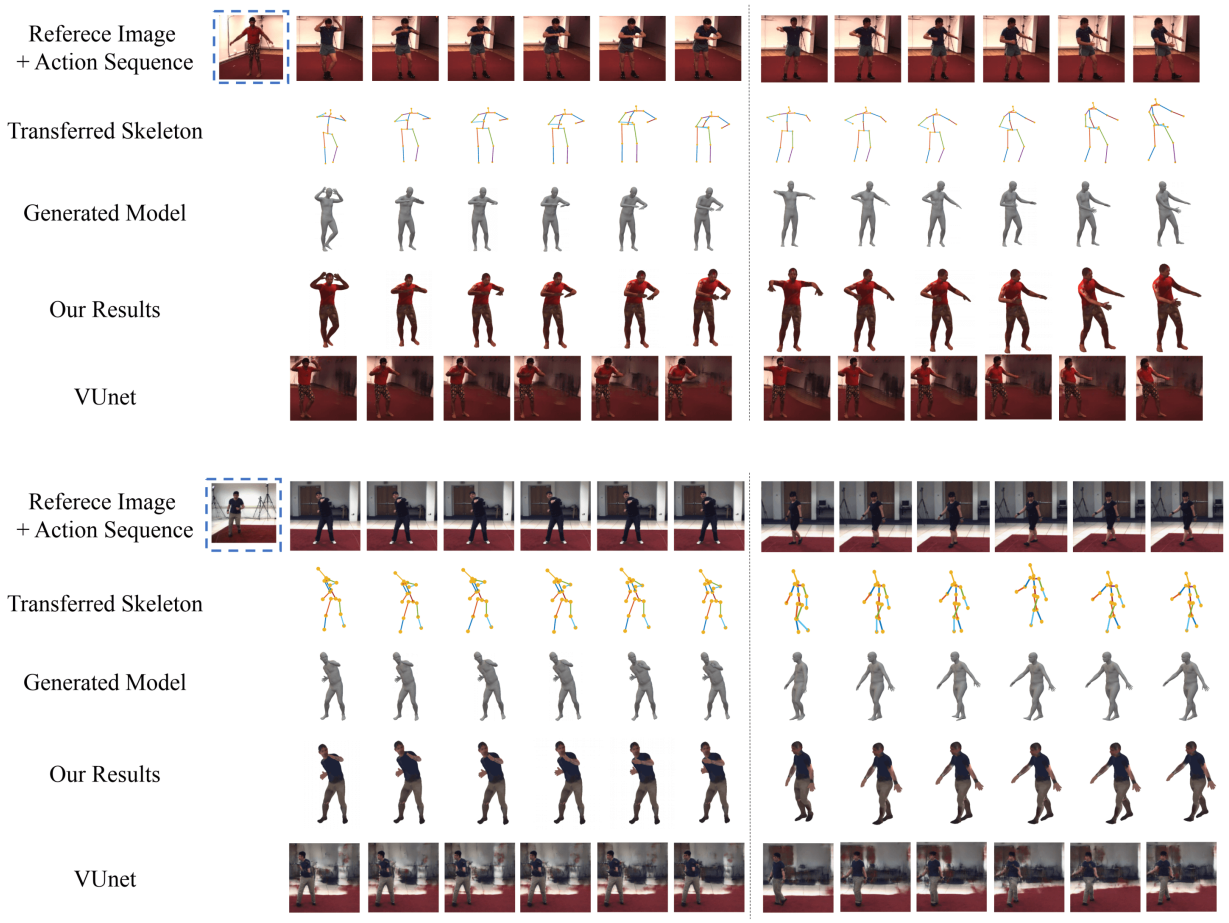

Figure 6: The experiment results of our framework on Human3.6M dataset (the first two lines) and HumanEva-I dataset (the last line). From top to bottom shows the target person and the source action video sequence, adapted skeleton sequence, generated human 3D model sequence, final transferred results, and the results from Variational U-net (Esser, Sutter, and Ommer 2018).

Transfer Results

We evaluate the whole proposed framework on both Human3.6M and HumanEva-I datasets. Moreover, there have been a few works that can perform human action transfer tasks following the deep learning methods. Here, we choose the most competitive work proposed in (Esser, Sutter, and Ommer 2018) as our baseline, which is able to learn from the appearance and synthesis images at different poses according to the given appearance. The comparative result is shown in Figure 6, from which we can see that our method have better performance on retention of texture details and body shape consistency.We also evaluate the mean value of SSIM and PSNR on both methods, and the result is shown in Table 1.

	Human3.6M		HumanEva-I	
	$\overline{SSIM}$	$\overline{PSNR}$	$\overline{SSIM}$	$\overline{PSNR}$
VUnet	0.88032	30.26613	0.87658	27.21103
Ours	0.90103	33.98061	0.89630	30.69898

Table 1: Transfer results comparison with Variational U-net. Mean SSIM and mean PSNR are used to assess the results

Rendering Results

For the texture rendering step, we evaluate it individually on the Human3.6M dataset and People-snapshot dataset. As mentioned before, the number of images of the target person of different views needed is hard to determine, as fewer images will provide insufficient appearance information while more images will increase information redundancy and lead to a conflict. Figure 7 shows that 3 images of different views is optimal for texture extraction by assessing the quality of texture image using Brenner gradient. In addition, we find our texture rendering method also has advantages when compared to the approach proposed in (Villegas et al. 2018), which is able to reconstruct 3D human model from a video of target person with fixed pose. For fair comparison, results of our method use the same model generated by (Alldieck et al. 2018). As shown in Figure 8, with the help of pixels' semantic information, we can see that our method is better when we use 3 images of different views.

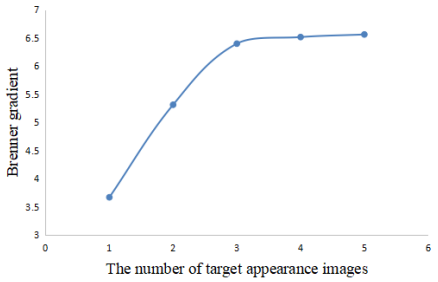

Figure 7: Mean Brenner gradient of the texture image vs. the number of appearance images in different views. When using 3 appearance images, the texture image is of the best quality.

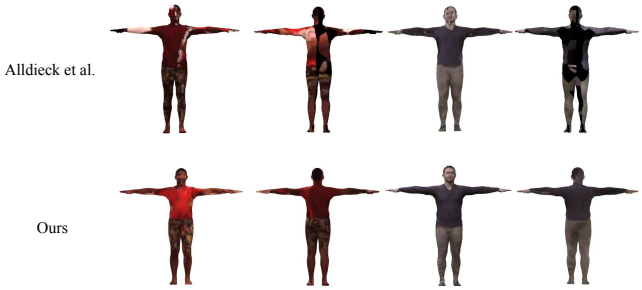

Figure 8: The rendering results comparison using 3 appearance images between our method and (Alldieck et al. 2018). Results of our methods are better when considering semantic information.

Ablation Study

Structure Comparison We analyze the effect of some individual components in our proposed framework on the final transfer performance. As shown in Figure 4, the base architecture of the network we use to produce SMPL pose parameter (θ) does not consider the influence from the previous pose. Therefore we improve this architecture to allow the network to take the information of previous poses into consideration. To verify that temporal structure can improve the smoothness of results, we use human evaluation metric. Firstly, we select 50 videos from test set of Human3.6M and HumanEva-I, each sequence contains 10 frames. Then we ask 100 volunteers to do MOS test on results, which are generated by basic model and temporal model respectively. Finally, we calculate the average fraction, shown in Table 2. We can find that temporal connection can produce better visual results.

Loss Analysis To verify the effectiveness of prior loss, we show the compare results on Human3.6M in Figure 9. Firstly, We randomly choose two images from Human3.6M and put their 3D skeletons into pose generator to get the 3D model. Given the 3D model, we reproject the skeleton into the images to compare the accuracy of joint locations between the model with prior loss and that without prior loss. Both reprojected skeleton and generated models are shown in Figure 9. Reprojected skeleton results indicate that the model without prior loss can fit correct joint locations, but the model with prior can fit more precisely. What's more, without prior loss, the model will predict twisted ankles compared to the model with prior loss.

	Human3.6M	HumanEva-I
Basic Model	3.89 ± 0.01	3.91 ± 0.06
Temporal Model	4.13 ± 0.09	4.20 ± 0.11

Table 2: Mean Opinion Score (MOS) of base pose generator model and the improved model with temporal connection. The higher MOS score indicates better results.

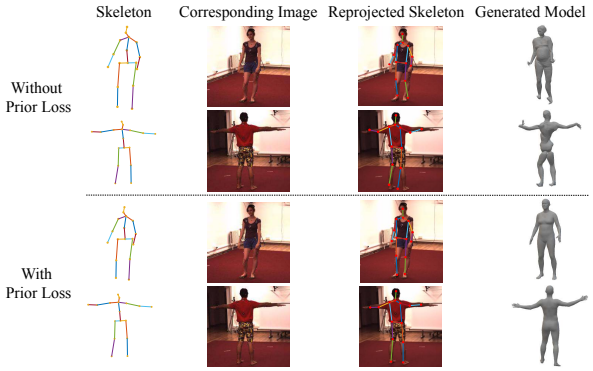

Figure 9: 3D human model generation results comparison between generators with and without prior loss. The generator without prior loss will generate unreasonable human shape.

Conclusion

To solve the upper bound of traditional 2D generation method, we propose a novel network to improve action transfer in 3D viewpoint. With stronger structure prior than GAN or VAE, our skeleton-to-3D-mesh generator is able to generate complete and time-smooth human shape. In addition, our texture generating process has more slack limitation to human pose than those in graphics.

Acknowledgements

The work was partly supported by State Key Research and Development Program (2016YFB1001003),National Science Foundation of China (U1611461, 61502301, 61527804, 61521062), China's Thousand Youth Talents Plan, STCSM (18DZ1112300, 17511105401, 18DZ2270700).

References

Akhter, I., and Black, M. J. 2015. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1446–1455.

Alldieck, T.; Magnor, M.; Xu, W.; Theobalt, C.; and Pons-Moll, G. 2018. Video based reconstruction of 3d people

models. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bellini, R.; Kleiman, Y.; and Cohen-Or, D. Dance to the beat: Enhancing dancing performance in video.

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

Brenner, J. F.; Dew, B. S.; Horton, J. B.; King, T.; Neurath, P. W.; and Selles, W. D. 1976. Automated microscope for cytologic research: Preliminary evaluation. *Journal of Histochemistry & Cytochemistry* 24(1):100–111.

Chan, C.; Ginosar, S.; Zhou, T.; and Efros, A. A. 2018. Everybody Dance Now. *ArXiv e-prints*.

Choi, K. J. 2000. Online motion retargeting. *Journal of Visualization & Computer Animation* 11(5):32–42.

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

Fan, L.; Huang, W.; Gan, C.; Ermon, S.; Gong, B.; and Huang, J. 2018. End-to-end learning of motion representation for video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gleicher, M. 1998. Retargetting motion to new characters. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '98, 33–42. ACM.

Güler, R. A.; Neverova, N.; and Kokkinos, I. 2018. Densepose: Dense human pose estimation in the wild. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hecker, C.; Raabe, B.; Enslow, R. W.; DeWeese, J.; Maynard, J.; and van Prooijen, K. 2008. Real-time motion retargeting to highly varied user-created morphologies. In *ACM SIGGRAPH 2008 Papers*, SIGGRAPH '08, 27:1–27:11. ACM.

Huang, Y. 2017. Towards accurate marker-less human shape and pose estimation over time. In *3D Vision (3DV), 2017 International Conference on*, 421–430. IEEE.

Ionescu, C.; Li, F.; and Sminchisescu, C. 2011. Latent structured models for human pose estimation. In *International Conference on Computer Vision*.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2014. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36(7):1325–1339.

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

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lassner, C.; Romero, J.; Kiefel, M.; Bogo, F.; Black, M. J.; and Gehler, P. V. 2017. Unite the people: Closing the loop between 3d and 2d human representations. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 3.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34(6):248:1–248:16.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch.

Pavlakos, G.; Zhu, L.; Zhou, X.; and Daniilidis, K. 2018. Learning to estimate 3d human pose and shape from a single color image. *arXiv preprint arXiv:1805.04092*.

Popa, A.-I.; Zanfir, M.; and Sminchisescu, C. 2017. Deep multitask architecture for integrated 2d and 3d human sensing. In *Conference on Computer Vision and Pattern Recognition*, volume 1, 5.

Sigal, L.; Balan, A. O.; and Black, M. J. 2009. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision* 87(1):4.

Sigurdsson, G. A.; Russakovsky, O.; and Gupta, A. 2017. What actions are needed for understanding human actions in videos? In *IEEE International Conference on Computer Vision*, 2156–2165.

Tak, S., and Ko, H.-S. 2005. A physically-based motion retargeting filter. *ACM Trans. Graph.* 24(1):98–117.

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

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13(4):600–612.

Yang, H.; He, X.; and Porikli, F. 2018. One-shot action localization by learning sequence matching network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zanfir, M.; Popa, A.-I.; Zanfir, A.; and Sminchisescu, C. 2018. Human appearance transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5391–5399.

Zhou, Y.; Sun, X.; Zha, Z.-J.; and Zeng, W. 2018. Mict: Mixed 3d/2d convolutional tube for human action recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.