

**METHOD AND SYSTEM FOR ONE-SHOT HUMAN MOTION TRANSFER
VIA OCCLUSION-ROBUST FLOW PREDICTION
AND NEURAL TEXTURING**

PRIORITY CLAIM

[0001] This application claims priority from Singapore Patent Application No. 10202250404H filed on 08 July 2022.

TECHNICAL FIELD

[0002] The present invention generally relates to avatar-based virtual environments, and more particularly relates to a method and a system for one-shot human motion transfer to avatars in virtual environments based on occlusion-robust flow prediction and neural texturing.

BACKGROUND OF THE DISCLOSURE

[0003] Social and work-related human interaction have been dramatically shifting online, partly driven by pandemic measures, but also supported by increasing bandwidth and the ease of digital connectivity. However, existing platforms, such as video conferencing and social media, has come with a loss of shared presence, when compared with in-person interactions.

[0004] Efforts to try and reclaim this sense of presence include the development of embodied avatar-based platforms where users experience a shared virtual world. This was first popularized by *Second Life*, a metaverse created and managed by Linden Labs of San Francisco, California, USA, while in the virtual gaming world there has been a continual evolution of multiplayer online video games. More recently, Meta Platforms, Inc. of Menlo Park, California, USA, generated heavy publicity for their metaverse *Horizon Worlds*, followed by other companies also launching similar ecosystems. Such

platforms are not only able to be experienced on normal 2D displays, but are well-suited for use with virtual reality headsets.

[0005] In such virtual worlds, users are embodied in the form of virtual avatars. There are however two challenges for greater adoption: visual appearance and motion transfer. First, the visual appearance of virtual avatars is seldom realistic, often appearing cartoon-like, possibly to avoid the well-known uncanny valley phenomenon. The uncanny valley phenomenon is a phenomenon in which a human being experiences a negative emotional response to a non-human figure, object, or entity that possesses a human likeness. The underlying idea is that humans are comfortable with human-like objects or entities that imperfectly resemble actual human beings up until a point that such humanoid objects provoke uncanny or strangely familiar feelings of uneasiness and revulsion in observers. While a human's affinity for such human-like objects or entities tends to increase with the entity's human likeness, the "valley" denotes the dip in the human observer's affinity.

[0006] The second challenge, and a harder one, is that it is difficult to transfer body, hand and head movements, including facial expressions and hand gestures, from users to avatars with a sufficient degree of richness and fidelity. In many instances, such motion transfer may not even be provided for such as when there are only basic movement controls via a keyboard or mouse input.

[0007] Thus, there is a need for methods and systems for motion transfer to avatars in a virtual environment which overcomes the drawbacks of current methods and systems and provides a more realistic motion of the avatar. Furthermore, other desirable features and characteristics will become apparent from the subsequent detailed description and the appended claims, taken in conjunction with the accompanying drawings and this background of the disclosure.

SUMMARY

[0008] According to at least one aspect of the present embodiments, a method for avatar animation in a virtual environment is provided. The method includes receiving first data of at least one source image corresponding to an avatar to be animated in the virtual environment and receiving second data corresponding to the desired animation. The method further includes combining the first data and the second data to animate the avatar by processing the first and second data to generate a motion-transferred image in response to both latent-feature warping of appearance variation between the first and second data and neural texture mapping of coarse spatial differences between the first and second data.

[0009] According to another aspect of the present embodiments, a system for avatar animation in a virtual environment is provided. The system for avatar animation receives first data of at least one source image corresponding to the avatar to be animated in the virtual environment and second data corresponding to the desired animation and includes a multi-scale feature warping module, a neural texture mapping module and a BlenderNet module. The multi-scale feature warping module is configured to latent-feature warp appearance variations between the first and second data and the neural texture mapping module is configured to neural texture map coarse spatial differences between the first and second data. The BlenderNet module is coupled to the multi-scale feature warping module and the neural texture mapping module and is configured to generate a motion-transferred image to animate the avatar in the virtual environment in response to both the latent-feature warp appearance variations between the first and second data and the neural texture map of coarse spatial differences between the first and second data.

[0010] According to yet a further aspect of the present embodiments, a non-transitory computer readable medium for acoustic sensing is provided. The non-transitory computer readable medium has stored thereon software instructions that, when executed by a processor, cause the processor to perform method that at least includes receiving first data of at least one source image corresponding to an avatar to be animated in the virtual environment, receiving second data corresponding to the desired animation, and combining the first data and the second data to animate the avatar by processing the first and second data to generate a motion-transferred image in response to both latent-feature warping of appearance variation between the first and second data and neural texture mapping of coarse spatial differences between the first and second data.

BRIEF DESCRIPTION OF THE DRAWINGS

[0011] The accompanying figures, where like reference numerals refer to identical or functionally similar elements throughout the separate views and which together with the detailed description below are incorporated in and form part of the specification, serve to illustrate various embodiments and to explain various principles and advantages in accordance with present embodiments.

[0012] FIG. 1, comprising FIGs. 1A, 1B and 1C, depicts exemplary avatar images from Horizon Worlds and Oculus Avatars.

[0013] FIG. 2 depicts an illustrative example of operation of methods and systems in accordance with the present embodiments.

[0014] FIG. 3 depicts a block diagram of the architecture of a motion transfer system for animation of a source image in accordance with the present embodiments.

[0015] FIG. 4, comprising FIGs. 4A to 4D, depicts block diagram of the modules of the motion transfer architecture in accordance with the present embodiments, wherein FIG. 4A depicts a block diagram of the MotionNet module, FIG. 4B depicts a block diagram of the multi-scale feature warping module, FIG. 4C depicts a block diagram of the neural texture mapping module, and FIG. 4D depicts a block diagram of the BlenderNet module.

[0016] FIG. 5, comprising FIGs. 5A to 5E, depicts images comparing motion transfer in accordance with the present embodiments with conventional methods, wherein FIG. 5A depicts the source image, FIG. 5B depicts the driving pose, FIG. 5C depicts a first first-order motion model (FOMM) motion transfer result, FIG. 5D depicts a second first-order motion model (FOMM-AA) motion transfer result, and FIG. 5E depicts motion transfer result in accordance with the present embodiments.

[0017] FIG. 6, comprising FIGs. 6A to 6D, depicts images comparing motion transfer in accordance with the present embodiments with conventional methods for source images and driving sequences from a TED Talks dataset.

[0018] FIG. 7, comprising FIGs. 7A and 7B, depicts images comparing motion transfer in accordance with the present embodiments with conventional methods for source images and driving sequences from a Tai Chi HD dataset.

[0019] FIG. 8, comprising FIGs. 8A and 8B, depicts images comparing motion transfer in accordance with the present embodiments with conventional methods for source images and driving sequences from an iPER dataset.

[0020] And FIG. 9 depicts images comparing motion transfer in accordance with the present embodiments with conventional methods for source images and driving sequences from a TED Talks dataset.

[0021] Skilled artisans will appreciate that elements in the block diagrams are illustrated for simplicity and clarity and have not necessarily been depicted to scale.

DETAILED DESCRIPTION

[0022] The following detailed description is merely exemplary in nature and is not intended to limit the invention or the application and uses of the invention. Furthermore, there is no intention to be bound by any theory presented in the preceding background of the invention or the following detailed description. It is the intent of present embodiments to present methods and systems for animating a static image of a person to fit the corresponding motion in a driving video sequence, while preserving the original person's identity. A unified framework is presented in accordance with present embodiments that combines multi-scale feature warping and neural texture mapping to better recover both two-dimensional (2D) appearance and 2.5D geometry. The methods and systems in accordance with the present embodiments provide a framework which, by joint training and multimodal fusion, learns robust neural texture features that cope with geometric errors as well as multi-scale dense motion flow to better preserve appearance of the transferred motion, providing a model that can generalize well and achieve advantageous motion transfer results which is particularly effective in handling challenging cases such as those with substantial self-occlusions

[0023] It is noted that learning to disentangle motion and identity is of central importance in achieving more realistic motion in a virtual environment. However, due to the diversity of motion patterns, different identities with different body shapes, and self-occlusion issues, it remains challenging to render photo-realistic human images based on specified driving motions and generating controllable human animations in both computer vision and graphics. Accordingly, many current avatars are rendered

with a cartoonish appearance and only provide limited motion control. FIGs. 1A, 1B and 1C provide examples of such existing avatar examples from Horizon Worlds and Oculus Avatars.

[0024] In recent years, Generative Adversarial Network (GAN) based models have been proposed for human motion transfer. GAN is a deep learning, unsupervised machine learning architecture which includes at least two main blocks: a generator and a discriminator. The generator tries to generate images which are very similar to that of an original dataset by taking noise as input and attempting to learn a joint probability ($P(X|Y)$) of the input data (X) and output data(Y). The discriminator tries to accept two inputs, one from a main dataset and another from images generated by the generator, and bifurcates them as 'Real' or 'Fake'. To make this generative and adversarial process simple, both the generator block and the discriminator block are made from deep neural network-based architecture which can be trained through forward and backward propagation techniques.

[0025] Conventional solutions treated motion transfer as an aligned image-to-image translation task. However, by leveraging a video-to-video (vid2vid) translation framework, models can learn to map the input of keypoints or skeletons to photorealistic human body images. Yet, solutions utilizing a vid2vid framework typically constrain motion transfer to a single person and are limited by generalization capabilities.

[0026] Alternatively, the problem can be treated as an unaligned image-to-image translation task by using conditional input such as body landmarks, articulated skeletons or semantic masks. The focus of such solutions is on establishing a correspondence between source and driving poses via attention mechanisms or correlation matrices.

[0027] One conventional model has been proposed using a first-order motion model (FOMM) based on self-supervised keypoints and two-dimensional affine regional

motion. This model was later extended to better disentangle shape and pose, as this was found to be more appropriate for articulated motion. Another model proposed to progressively transfer a conditional pose through a sequence of intermediate pose representations by leveraging given source and driving landmarks. And yet another solution used a differentiable global-flow local-attention framework to spatially transform inputs at the feature level.

[0028] Pose-guided image generation methods aim to transfer pose for arbitrary persons, but will not perform well for animation without temporal modelling. Moreover, such existing methods tend to work well only for images within the same dataset, which leads to limited generalization ability.

[0029] A major challenge in motion transfer methods is finding correspondence between source and driving poses due to self-occlusion and missing body parts. Artefacts from unsupervised methods are more significant due to semantic ambiguity in identifying body parts without supervision. In addition, unsupervised 2D-based methods may get confused by whether a person is facing front or back as 2D landmarks, regions, or masks may not provide adequate guidance of body pose.

[0030] While recent advances in one-shot human motion transfer have led to significant improvement in results, it remains difficult to capture accurate correspondences between source and driving poses due to the large variation in motion and articulation complexity.

[0031] The human motion transfer framework of the methods and systems of the present embodiments addresses animating a static image of a person to fit the corresponding motion in a driving video sequence, while preserving the original person's identity. Thus, disentangling the motion and identity is of central importance. However, due to the diversity of motion patterns, different identities with different body

shapes, and self-occlusion issues, it is challenging to render photo-realistic human images based on specified driving motions and generating controllable human animations in both computer vision and graphics.

[0032] Recent advances in motion transfer have adopted two main paradigms: pose-guided image/video generation and neural rendering with neural texture features. Specifically, pose-guided image/video translation methods directly warp appearance features through various correspondences between source and driving pose, including through dense motion flow, attention and cross-domain correlations. These methods can achieve good performance in preserving identity but do not handle large variations in body pose well. A key issue is the self-occlusion within the motion that corrupts the accuracy of correspondence for feature alignment resulting in reduced quality of synthesized images. In addition, such methods may be confused about whether a person is facing front or back when using degenerate 2D supervision, such as using 2D landmarks, articulated structures and semantic masks.

[0033] Separately, neural-rendering-based methods can achieve high-fidelity and photorealistic human image generation and animation. However, most of these models need to be trained in a person-specific fashion and cannot easily be generalized to new individuals. Recently, efforts have been made to generalize neural-rendering-based methods to arbitrary persons by utilizing auxiliary 2.5D or 3D prior information, such as a DensePose IUUV map and/or a Skinned Multi-Person Linear (SMPL) model. DensePose is a real-time approach for mapping all human pixels of 2D RGB images to a 3D surface-based model of the body from Meta Platforms, Inc. that establishes dense correspondences in an IUUV map from a 2D image (an inferred map having components “I”, “U” and “V”) to a 3D surface-based representation of the human body (having the coordinates “x”, “y” and “z”). A SMPL model is a realistic 3D model of the human

body that is based on skinning and blend shapes and is learned from thousands of 3D body scans. However, the auxiliary structure information in the DensePose IUV map and SMPL model can be noisy, which results in a significant loss of 2D appearance information. In particular, a DensePose IUV map occasionally suffers from large gaps, while a SMPL model only models the human body without hair and clothing.

[0034] As described hereinbelow, the methods and systems in accordance with the present embodiments present an improved visually realistic experience by looking to some aspects of DensePose and SMPL modelling for guidance. In particular, despite its limited accuracy and restriction to the SMPL silhouette, DensePose can be used as guidance in two ways. First, by combining with the appearance context of a source image, DensePose can be used to predict a more accurate dense motion flow for feature warping, even for regions that are not originally covered by DensePose. And secondly, DensePose can be used to support texture atlas translation and mapping, via neural textures that can overcome inaccuracies in coarse geometry. Furthermore, the joint training of 2.5D geometry and 2D appearance features may correct for geometrical errors and stabilize the generation of images through multi-modal feature fusion. Thus, it can be said that, while DensePose is able to provide coarse but occlusion-aware 2D correspondences, the methods and systems in accordance with the present embodiments utilizes such 2D correspondences of DensePose partly by exploiting the information from DensePose, yet adapting to its inherent limited accuracy.

[0035] Accordingly, to address the drawbacks of previous systems and the issues discussed hereinabove, the methods and systems in accordance with the present embodiments focuses on animating an arbitrary source photograph of a human subject (i.e., a “source” person S) with an observed motion from a different person (i.e., a “driving” person D). The broad objectives of the methods and systems in accordance

with the present embodiments are to preserve as much as possible of the visual realism of S, while retaining as much as possible the fidelity of D's motion. In particular, these objectives are addressed in a one-shot setting in which only one source image of S is provided at a time of use (i.e., during inference or test time), and S can be previously unseen (i.e., images of the person S are not the training dataset).

[0036] More concretely, three steps are to be taken by a user to animate a source image in accordance with the present embodiments. First, the user takes a photo of a person, or selects an image of a person from the internet or some existing dataset, the phot being a source image. The person in the source image can be in any pose. Next, the user stands in front of a camera and either acts naturally or otherwise performs some intended action. Then as a third and final step, the methods and systems in accordance with the present embodiments takes the video from the camera and animates the subject in the source image by transferring the motion from the user to the subject. An illustration 200 of FIG. 2 depicts a broad outline of these three steps. The photos 210 are a variety of source images with different identities, clothing and poses which can be selected for animation. Only one source image is sufficient and the user can select from images such as those pictured or take a photo of a person.

[0037] At the second step 220, the user stands in front of a video camera and performs various actions. The results of the methods and systems in accordance with the present embodiments are shown in the images 230 where the source images (i.e., the images 210) are independently animated with the user's motion. It is clear from the images 230 that the animation visual quality is realistic rather than cartoonish and the pose of the driving person has been reasonably well transferred.

[0038] As an overview to the methods and systems in accordance with the present embodiments, given a static image $\mathbf{I}_s$ and the extracted dense pose $\mathbf{P}_s$, the main goal of

the architecture is to learn a model Φ that synthesizes a photorealistic human video sequence $\{\hat{\mathbf{I}}_d^{(1)}, \hat{\mathbf{I}}_d^{(2)}, \dots, \hat{\mathbf{I}}_d^{(N)}\}$, in a manner that retains the identity of the source image $\mathbf{I}_s$ but appropriately drives its pose as the dense pose sequence $\{\hat{\mathbf{P}}_d^{(1)}, \hat{\mathbf{P}}_d^{(2)}, \dots, \hat{\mathbf{P}}_d^{(N)}\}$ extracted from the corresponding input video.

[0039] Referring to FIG. 3, a block diagram 300 depicts the modular architecture in accordance with the present embodiments which combines the learning of appearance and geometry into a multi-stream pipeline. This framework combines the advantage of pose-guided and neural-rendering-based methods in order to learn to synthesize images with better recovery of appearance and geometry by integrating the multi-modal features into the modular architecture. Four modules of the modular architecture pictured in the block diagram 300 are a MotionNet module 310, a multi-scale feature warping module 320, a neural texture mapping module 330, and a BlenderNet module 340.

[0040] The MotionNet module 310 utilizes a motion network for producing dense motion flow and translation signals for both appearance and geometry translation. The motion network is presented to simultaneously predict the dense flow for 2D appearance feature warping and the translation signals for the 2.5D dense pose branch.

[0041] The multi-scale feature warping module 320 utilizes a feature warping network for accurate appearance feature alignment in the 2D image plane. The neural texture mapping module 330 recovers and stabilizes the DensePose annotation by joint training a neural texture atlas translation and a DensePose translation. The feature warping network of the multi-scale feature warping module 320 is applied to capture multi-scale identity features with respect to different scales, while separately the neural texture mapping module 330 is designed to achieve arbitrary neural body image rendering by disentangling the texture and geometry.

[0042] Finally, the BlenderNet module 340 provides image refinement and super-resolution to further improve the translated image quality by fusing the translated appearance and geometry features.

[0043] Experimental results show that the architecture in accordance with the present embodiments can benefit from both appearance and geometry information, achieving competitive performance on animating images with the correct pose. Challenging motion patterns, including full body rotation, self-occlusion and subtle hand motions, can be transferred well. The unified architecture in accordance with the present embodiments combines feature warping and neural texture mapping to animate the human body images with high-quality pose transfer. In addition, the architecture implicitly disentangles neural texture atlas translation and DensePose translation for better recovery and stabilization of the DensePose geometry under a differentiable neural texture mapping framework. Results on video datasets with half and full human body demonstrate that the architecture in accordance with the present embodiments can be applied for challenging motion transfer cases such as turning around, self-occlusion, and front-to-back view pose transfer.

[0044] Three-dimensional geometry representations for pose transfer can be more effective for larger pose variations, especially for handling occlusion and de-occlusion. Some recent neural rendering-based methods made use of DensePose and parametric 3D models to improve synthesis quality. Core ideas, such as disentangling the representation of human geometry and texture, inpainting a complete set of UV texture from visible parts, few-shot learning for texture atlas completion, and 3D-based fine-grain appearance flow have been proposed to achieve better motion transfer.

[0045] One solution presented a two-stream pipeline of pose transfer by combining surface-based image synthesis and texture atlas warping. Inspired by the two-stream

pipeline approach, a second solution proposed a neural re-rendering approach to learn a high-dimensional detail-preserving UV feature map, while a third solution presented a few-shot motion transfer method that aimed to learn complete texture atlases.

[0046] The neural texture mapping module 330 is closely related to neural re-rendering approach and the few-shot motion transfer method, but the neural texture mapping module 330 utilizes neural texture translation that learns texture maps conditioned on each driving DensePose rather than attempting the ill-posed task of learning a complete texture atlas. In addition, due to typical DensePose errors, the neural texture mapping module 330 does not attempt to extract texture maps in a separate data preparation stage.

[0047] Unlike existing methods that directly predict the geometry and appearance from input pose and image, the MotionNet module 310 attempts to estimate the residual information of appearance and geometry by accounting for the difference between the target pose $\mathbf{P}_d$ and the source pose $\mathbf{P}_s$, in the context of the source image $\mathbf{I}_s$. Instead of directly driving the person in a source image, it is easier to learn the residual difference between the input pose and target pose. In particular, the MotionNet module 310 first estimates a pixel-level dense motion flow for high-quality visual appearance warping in the feature warping module 320 and also learns feature-level translation signals α and ρ for the disentangled texture atlas translation and dense pose translation used in the neural texture mapping module 330. One notable aspect is that while $\mathbf{P}_d$ and $\mathbf{P}_s$ establish explicit correspondence between some pixels (though with limited accuracy), these are limited to pixels within the DensePose silhouettes. The MotionNet module 310 learns to exploit this strong guidance to also predict motion flow for pixels outside the silhouettes, such as for hair and clothing, in relation to image features from $\mathbf{I}_s$.

[0048] Referring to FIG. 4A, a block diagram of the MotionNet module 310 is depicted. The structure of the MotionNet module 310 has a general encoder/decoder architecture. Specifically, the encoder 400 is responsible for extracting feature-level translation signals for the 2.5D geometry branch, while the decoder 410 has an iteratively upsampled refinement framework to predict multiscale pixel-level dense motion flows and occlusion maps. This can be expressed as shown in Equation (1)

$$\left[\alpha, \rho, [\mathbf{f}^{(i)}, \mathbf{o}^{(i)}] \right] = \mathbf{MotionNet}(\mathbf{I}_s, \mathbf{P}_s, \mathbf{P}_d) \quad (1)$$

where α and ρ respectively provide guidance for neural texture atlas translation and dense pose translation, while $\mathbf{f}^{(i)}$ and $\mathbf{o}^{(i)}$ respectively denote the predicted i -th level dense motion flow and occlusion map.

[0049] In the encoder 400, a global average pooling (GAP) layer 402 is provided to obtain a latent representation of motion residual. Then, two separate fully connected layers (FC_1 404 and FC_2 406) are adopted to obtain translation signals for neural texture atlas and DensePose. The decoder 410 includes an iteratively upsampled refinement module to produce multi-scale dense motion flow. The lines 412, 414 denote the prediction heads of dense motion flows and occlusion maps.

[0050] Referring to FIG. 4B, a block diagram of the multi-scale feature warping module 320 is depicted. The MotionNet module 310 produces the multi-scale dense motion flow $\mathbf{f}$ and occlusion map $\mathbf{o}$ as initial transfer signals. In order to predict the more refined 2D transferred appearance in accordance with the present embodiments, it is necessary to align the source image content based on the predicted multi-scale motion flow and occlusion map. This function is implemented using a multi-level feature framework as shown in the multi-scale feature warping module 320 of FIG. 4B which can capture different scales of motion on different feature maps. More specifically, a given source image $\mathbf{I}_s$ is embedded in multi-scale features as shown in Equation (2)

$$\left[\mathbf{z}_s^{(1)}, \mathbf{z}_s^{(2)}, \dots, \mathbf{z}_s^{(L)} \right] = \mathcal{E}_s(\mathbf{I}_s) \quad (2)$$

where $\mathbf{z}_s^{(1)}$ denotes the i -th level source feature embedded by the encoder $\mathcal{E}_s$. In conjunction with the multi-scale motion flow $\mathbf{f}$ obtained from MotionNet module 310, the source feature is then warped through the residual motion flow. Furthermore, the additional occlusion maps $\mathbf{o}$ are used as filters to propagate the source information to the transferred feature: thus, when the occlusion is 1, the source features will be directly passed on, otherwise the previously warped features will be passed on. Formally, the operation is defined as shown in Equation (3)

$$\hat{\mathbf{z}}_s^{(i+1)} = \mathbf{Warp}(\mathbf{z}_s^{(i)}, \mathbf{f}^{(i)}) * \mathbf{o}^{(i)} + \hat{\mathbf{z}}_s^{(i)} * (\mathbf{1} - \mathbf{o}^{(i)}) \quad (3)$$

where Warp is the flow warping function which in practice can be implemented by a bilinear sampling operation, while the occlusion map $\mathbf{o}^{(i)}$ determines how the warped source feature should be combined with the current aligned feature $\hat{\mathbf{z}}_s^{(i)}$. The final fused appearance feature $\mathbf{z}_{app}$ is used to generate a coarse image. In addition, the multi-scale feature warping module 320 also predicts a human segmentation mask that enables extra supervision during training with a prediction head in the generator. The final output of the 2D feature warping branch is shown in Equation (4)

$$[\hat{\mathbf{I}}_{app}, \hat{\mathbf{m}}_{app}] = \mathcal{G}_{app}(\mathbf{z}_{app}) \quad (4)$$

where $\hat{\mathbf{I}}_{app}$ is the generated image and $\hat{\mathbf{m}}_{app}$ denotes the predicted binary mask.

[0051] The multi-scale feature warping module 320 is designed in accordance with the present embodiments to capture a more accurate appearance feature. FIG. 4B illustrates the specific structure of the network of the multi-scale feature warping module 320 and shows that the warped source feature will be partially propagated to the decoder part for preserving source appearance. To stabilize the estimation of motion flows and occlusion maps with respect to different scales, flow warping is further applied on the

source and generated image by using multi-scale dense motion flows and occlusion maps.

[0052] At the core of the architecture pipeline in accordance with the present embodiments is the neural texture mapping module 330 as shown in detail in FIG. 4C. The neural texture mapping module 330 aims to implicitly disentangle the appearance and geometry in 2.5D UV space. To enable this, a two-stream pipeline 420, 430 is provided as shown in the detailed network structure of the neural texture mapping module 330. More specifically, the neural texture mapping module 330 module consists of two main branches, i.e., the neural texture atlas translation 420 and the DensePose translation branch 430. The translated neural texture atlas and DensePose are utilized to produce the neural texture feature by a differentiable texture mapping layer 435 and the neural texture feature is rendered to an image 440.

[0053] The methods and systems in accordance with the present embodiments learns to implicitly disentangle the texture mapping with both appearance and geometry disentanglement in the neural texture atlas (appearance) translation branch 320 and the DensePose IUUV map (geometry) translation branch 330. Note that conventional neural rendering-based methods are bypassed by the methods and systems in accordance with the present embodiments with three unique considerations. First, the existing neural texture mapping methods decompose the correspondence of appearance and geometry using DensePose, which is based on the SMPL model and strictly extracts the pixels within the SMPL silhouette ignoring other parts, such as hairs and background. Second, errors in the extracted DensePose IUUV map, including occasional large missing gaps, have a significant negative effect on the corresponding extracted textures. And third, direct extraction of texture atlases or part-based normalized textures based on DensePose correspondence may not only be time-consuming (unless carefully GPU-optimized), but

the quality of the extracted texture atlases may also be poor due to the limited resolution accuracy of DensePose.

[0054] The two-stream neural texture mapping framework in accordance with the present embodiments tames the model to implicitly decouple the translation of the neural texture atlas (i.e., by the branch 320) and the translation of the DensePose IUUV map (i.e., by the branch 330). In particular, given a source image $\mathbf{I}_s$ as well as the corresponding extracted DensePose map $\mathbf{P}_s$ and translation signals α, ρ (extracted by the MotionNet module 310), the texture atlas and geometry translation are respectively formulated as shown in Equation (5).

$$\hat{\mathbf{T}}_d = \mathcal{T}_{app}(\mathcal{E}_{app}(\mathbf{I}_s, \mathbf{P}_s), \alpha), \hat{\mathbf{P}}_d = \mathcal{T}_{geo}(\mathcal{E}_{geo}(\mathbf{I}_s, \mathbf{P}_s), \rho) \quad (5)$$

where $\mathcal{E}_{app}$ and $\mathcal{E}_{geo}$ are the appearance encoder and geometry encoder, respectively. They have some initial shared weights, but with two separate convolutional layers. $\mathcal{T}_{app}$ and $\mathcal{T}_{geo}$ refer to translation networks that output the neural texture atlas $\hat{\mathbf{T}}_d \in \mathbb{R}^{24 \times c \times h \times w}$ and the translated DensePose IUUV map $\hat{\mathbf{P}}_d$, respectively. In detail, the translated DensePose $\hat{\mathbf{P}}_d$ consists of three components, i.e. $[\hat{\mathbf{U}}, \hat{\mathbf{V}}, \hat{\mathbf{S}}]$, where $\hat{\mathbf{U}} \in \mathbb{R}^{24 \times h \times w}$ and $\hat{\mathbf{V}} \in \mathbb{R}^{24 \times h \times w}$ are the predicted UV coordinates for twenty-four non-overlapping body parts defined in DensePose with spatial size $h \times w$, while $\hat{\mathbf{S}} \in \mathbb{R}^{25 \times h \times w}$ represents the semantic score map for the different parts. In practice, a softmax function is applied to estimate the probabilities of each pixel belonging to different body parts.

[0055] Once the translated neural texture atlas feature and the soft mask for each body part is predicted, a differentiable texture mapping module is used to produce the target neural texture feature, which is subsequently rendered to a coarse output. Equation (6) computes the neural texture feature assigned to the $[i, j]$ location

$$\hat{\mathbf{R}}_{ij}^{(k)} = \hat{\mathbf{T}}_d^{(k)} \left(\hat{\mathbf{U}}_{ij}^{(k)}, \hat{\mathbf{V}}_{ij}^{(k)} \right) \quad (6)$$

where i and j index to the predicted UV coordinates for the k -th body part. Once the UV coordinates are obtained, the corresponding neural texture feature is assigned to the $[i, j]$ location. In practice, the above process can be efficiently implemented using a differentiable bilinear sampling operation in which the k -th component of the UV coordinates will be regarded as a sampling grid and the k -th neural texture atlas feature is the input for sampling. After that, neural texture feature maps for different body parts $\hat{\mathbf{R}}^{(k)}$ are weighted summed according to the probability score map $\hat{\mathbf{S}}$ as shown in Equation (7)

$$\mathbf{z}_{geo} = \sum_{k=1}^{24} \hat{\mathbf{S}}^{(k)} \hat{\mathbf{R}}^{(k)} \quad (7)$$

where $\mathbf{z}_{geo} \in \mathbb{R}^{c \times h \times w}$ is the unified neural texture feature. In accordance with the present embodiments, the 0-th channel with background probability in $\hat{\mathbf{S}}$ is ignored in this differentiable texture mapping process.

[0056] Although the neural texture mapping and DensePose IUUV map are implicitly decoupled in a differentiable module, the ground truth for both parts have not been obtained. Hence, an important issue to address is how can the 2D image be leveraged to guide the learning of this neural texture mapping. This is tackled in accordance with the present embodiments by first rendering a coarse image using the neural texture feature map with a single convolutional layer. Also, a foreground human body mask is introduced as extra supervision to guide the training to improve UV completion and extraction of neural texture features. In accordance with the present embodiments, two separate prediction heads are adopted to generate the coarse image and predict the foreground mask as shown in Equation (8)

$$[\hat{\mathbf{I}}_{geo}, \hat{\mathbf{m}}_{geo}] = \mathcal{G}(\mathbf{z}_{geo}) \quad (8)$$

where $\hat{\mathbf{I}}_{geo}$ is the rendered coarse image and $\hat{\mathbf{m}}_{geo}$ is the predicted binary mask.

[0057] In the BlenderNet module 340, as shown in detail in FIG. 4D, a shallow blending network is applied to further improve the visual quality of transferred results by fusing features from the 2D appearance flow warping translation branch 320 and the 2.5D geometry neural texture mapping translation of the neural texture mapping module 330. In particular, the final translated RGB image $\hat{\mathbf{I}}_d$ and the corresponding binary mask $\hat{\mathbf{m}}$ are simultaneously predicted as shown in Equation (9)

$$[\hat{\mathbf{I}}_d, \hat{\mathbf{m}}] = \mathbf{BlenderNet}(\mathbf{z}_{geo}, \mathbf{z}_{app}, \hat{\mathbf{I}}_{app}) \quad (9)$$

[0058] In order for the BlenderNet module 340 to integrate the 2D flow-warping feature and 2.5D neural texture feature for better recovering appearance and geometry, a shallow U-Net structure 450 with three scales of downsampling blocks is adopted. To generate images with higher resolution, an upsampling and convolutional layers as a super-resolution block 455 is provided.

[0059] Multiple loss functions are used in training in the pipeline in accordance with the present embodiments. As discussed, the methods and systems in accordance with the present embodiments consists of four different branches. In the MotionNet module 310, perceptual correctness loss and an affine transformation regularization term are used to penalize the error dense flow field and to stabilize the dense motion estimation module in the 2D branch. In addition, a multi-scale consistency loss is introduced to preserve the feature warping consistency among different scales. In addition, a total-variation (TV) regularization term has also been introduced to smooth the estimated dense motion flow in each level of output of the MotionNet module 310. This is shown in Equation (10)

$$\mathcal{L}_{\text{motion}} = \lambda_{cor}\mathcal{L}_{correctness} + \lambda_{reg}\mathcal{L}_{regularization} + \lambda_{tv}\mathcal{L}_{TV} + \lambda_{con}\mathcal{L}_{consistency} \quad (10)$$

For TV loss and consistency loss, the TV loss in each scale of dense motion flow is calculated to regularize the model for producing smooth flow. For multiscale consistency loss in accordance with the present embodiments, the flow with the lowest resolution (i.e., 32x32 in our implementation) is used as soft label to calculate the consistency loss with flows with higher resolutions (i.e., 64x64 and 128x128 in the present embodiments). In the experiments described hereinbelow, the weight hyperparameters λ_{cor} , λ_{reg} , λ_{tv} and λ_{con} are set to 5.0, 0.01, 1.0 and 5.0, respectively.

[0060] In 2.5D branch of the neural texture mapping module 330, cross-entropy (CE) loss is adopted for the semantic body part segmentation (i.e., the semantic mask prediction of the DensePose IUUV map) and a masked L1 loss for UV coordinate regression. For the neural rendered RGB image, since the differentiable neural texture mapping is based on the correspondence between the neural texture atlas and UV coordinates, pixels outside the predicted DensePose silhouettes should not be considered for calculating L1 reconstruction loss. Thus, a masked L1 loss is applied to filter out the regions outside the human body as shown in Equation (11)

$$\begin{aligned} \mathcal{L}_{iuv} = & \lambda_{uv} \sum_{k=1}^{24} (\|\hat{\mathbf{S}}^{(k)} \odot (\mathbf{U}^{(k)} - \hat{\mathbf{U}}^{(k)})\| + \|\hat{\mathbf{S}}^{(k)} \odot (\mathbf{V}^{(k)} - \hat{\mathbf{V}}^{(k)})\|) \\ & + \lambda_{ce} \mathcal{L}_{CE}(\mathbf{S}^{(k)}, \hat{\mathbf{S}}^{(k)}) \end{aligned} \quad (11)$$

where $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ denote the predicted UV components of the twenty-four human body parts, $\hat{\mathbf{S}}$ represents the probability score map, and $[\mathbf{U}, \mathbf{V}, \mathbf{S}]$ are the pseudo ground-truths provided by the DensePose IUUV map. In the present embodiments, λ_{uv} and λ_{ce} are set to 5.0, and 1.0, respectively.

[0061] For the feature warping module 320 and blending networks of the BlenderNet module 340, an image pyramid perceptual loss and L1 reconstruction loss are introduced to obtain better image quality. In accordance with the present embodiments, the perceptual loss and reconstruction loss are adopted for the generated RGB image of the

2D branch of the feature warping module 320 and the blending networks of the BlenderNet module 340 as shown in Equation (12)

$$\begin{aligned}\mathcal{L}_{rec} &= \lambda_p \mathcal{L}_p(\mathbf{I}_d, \mathbf{I}_q) + \lambda_1 \mathcal{L}_1 \\ &= \lambda_p \sum_{l=1}^L \sum_{i=1}^N \left\| \Phi^i(\boldsymbol{\Omega}^l(\mathbf{I}_d)) - \Phi^i(\boldsymbol{\Omega}^l(\hat{\mathbf{I}}_q)) \right\|_2 + \lambda_1 \left\| \mathbf{I} - \hat{\mathbf{I}}_q \right\|_1\end{aligned}\quad (12)$$

where $\mathbf{q} = \{\mathbf{app}, \mathbf{geo}, \mathbf{d}\}$ denotes a set of images generated by different modules, $\mathbf{I}_d$ is the ground-truth image, $\boldsymbol{\Omega}^l$ is the l -th scale of image resized by using a down-sampling operation, and Φ^i is the i -th layer of the VGG-19 pretrained network. In the present embodiments, λ_p and λ_1 are set to 10.0, and 1.0, respectively.

[0062] A human foreground mask prediction head is adopted as extra guidance for the 2D, 2.5D and blending modules. Concretely, a Binary Cross-Entropy (BCE) loss in the 2D branch can not only be used to penalize the error in the estimated motion flow, but also encourage the model to recover the appearance in the human regions. For the neural texture mapping module 330, the human foreground mask can provide extra supervision to guide the generation of content outside the DensePose silhouettes but inside the boundary of the human foreground mask as shown in Equation (13)

$$\mathcal{L}_{mask}(\mathbf{m}, \hat{\mathbf{m}}) = \text{BCE}(\mathbf{m}, \hat{\mathbf{m}}_q) \quad (13)$$

where $\mathbf{m}$ is the pseudo ground-truth mask, and $\hat{\mathbf{m}}_q$ the human mask predicted by the different modules.

[0063] Finally, the Least-squares GAN (LSGAN) loss is utilized to train the whole adversarial framework to further improve the quality of generated images as shown in Equation (14)

$$\begin{aligned}\mathcal{L}_{adv}(\mathbf{D}, \mathbf{G}) &= \frac{1}{2} \mathbb{E}[(\mathbf{D}(\mathbf{I}_d) - 1)^2] + \frac{1}{2} \mathbb{E}[\mathbf{D}(\mathbf{G}(\mathbf{I}_s, \mathbf{P}_s, \mathbf{P}_d))^2] \\ &\quad + \frac{1}{2} \mathbb{E}[\mathbf{D}((\mathbf{G}(\mathbf{I}_s, \mathbf{P}_s, \mathbf{P}_d)) - 1)^2]\end{aligned}\quad (14)$$

where D denotes the Discriminator. In the present embodiments, a simple Discriminator with ResNet blocks is utilized.

[0064] In order to evaluate the methods and systems in accordance with the present embodiments, a pipeline model in accordance with the present embodiments was trained on three publicly available video benchmark datasets for human motion transfer and the results were evaluated. The three video benchmark datasets included a dataset of TED Talks, a high-definition data set of Tai Chi exercising (hereafter referred to as the “Tai Chi HD dataset”), and an Impersonator dataset (hereafter referred to as the “iPER dataset”). The TED Talks dataset contains 1,123 video clips for training and 131 clips for testing, the video clips segmented from TED Talks YouTube videos. In each video, the upper part of the speaker is cropped from the video according to annotated bounding boxes. The Tai Chi HD dataset consists of 2,927 video clips for training and 253 clips for testing, each video capturing full-bodied tai chi actions. The iPER dataset consists of 164 video clips for training and 42 video clips for testing. The video sequences in the iPER dataset are relatively longer than the videos in the TED Talks dataset and the Tai Chi HD dataset, and include A-poses and random actions. In the experiments, the video frames in each dataset were cropped and resized to 256×256. To obtain pseudo data annotation, the latest released DensePose model was used to get the IUUV map for video frames in the above three datasets. The model uses ResNet101 as the backbone and a DeepLabV3 prediction head. Moreover, a video human matting method was adopted to obtain the human foreground masks as pseudo ground-truth for training.

[0065] To stabilize the training process, the model in accordance with the present embodiments is trained with a multi-stage setting. Specifically, the geometry branch 430 of the neural texture mapping module 330 and the encoder 400 of the MotionNet module 310 are trained first. Then, the MotionNet module 310 and the multi-scale

feature warping module 320 are trained in a second stage after freezing the other parts of the model. Finally, the full model as shown in FIG. 3 is trained end-to-end.

[0066] During training, the model in accordance with the present embodiments is optimized by using an optimizer algorithm for first-order gradient-based optimization of stochastic objective functions, such as the Adam optimizer which is based on adaptive estimates of lower-order moments with a learning rate of $2e-4$ and proposed by Diederik Kingma and Jimmy Ba in their 2015 paper entitled “Adam: A Method for Stochastic Optimization”. For each dataset, the model was trained with the same hyper-parameter settings where the lengths of latent feature vectors of the texture atlas translation signal and the DensePose translation signal are set to 384 and 256, respectively. The number of training iterations was fixed to 600,000. In the first training stage, the geometry branch 430 of the neural texture mapping module 330 was trained with 100,000 iterations. Then, the multi-scale feature warping module 320 was trained with another 100,000 iterations. Thereafter, the full model was trained with 200,000 iterations end-to-end and the BlenderNet module 340 was fine-tuned with a super-resolution block with 200,000 iterations. The models for ablation analysis were trained by setting the batch size to 8, and the input and output resolutions were set to 128×128 and 256×256 , respectively. All of the models for comparison with state-of-the-art and ablation studies were trained and tested on a workstation with a GeForce RTX 3090Ti (24G) graphics processing unit (GPU) from NVIDIA of Santa Clara, California, USA.

[0067] In comparison to person-specific models, multi-camera models and two first-order motion models (FOMM and FOMM-AA), motion transfer and avatar animation by methods and systems in accordance with the present embodiments provide an undemanding, scalable solution. For example, as compared to person-specific models, while the methods and systems in accordance with the present embodiments are

technically harder to create high quality visual appearance, the approach in accordance with the present embodiments is more easily scalable and avoids the extensive manual labour needed to train person-specific models. The person-specific models require collecting extensive footage of the source individuals to be animated and undertaking a computationally costly training of deep networks for each source individual. Many of such person-specific methods also require multi-camera systems.

[0068] As compared to multi-camera systems, including generalizable and “few-shot” methods, the methods and systems in accordance with the present embodiments are simpler and easier to embrace. Multi-camera systems entail deploying a multi-camera setup in a clean/green-screened environment, which is hard for a general user to set up, thereby preventing wide adoption.

[0069] As compared to “one-shot” FOMM and FOMM-AA methods, the methods and systems in accordance with the present embodiments provide visually perceptible improvements. While the one-shot FOMM and FOMM-AA methods are general enough that they are not limited to human body motion transfer, the frameworks are established entirely in 2D motion with some occlusion map prediction, but not to the level of occlusion reasoning that is established in the methods and systems in accordance with the present embodiments. As a result, these methods have a difficult time coping with human motion that have substantial occlusion and de-occlusion. In particular, they may completely confuse the front-back orientation of driving users who are turning around on-the-spot. This can be seen in FIGs. 5A to 5E where FIG. 5A depicts the source image and FIG. 5B depicts the driving pose. FIG. 5C depicts the FOMM motion transfer result, FIG. 5D depicts the FOMM-AA motion transfer result, and FIG. 5E depicts the motion transfer result in accordance with the present embodiments. It can be seen that the purely 2D-based FOMM and FOMM-AA

methods get confused by the person's orientation. For example, in FIGs. 5C and 5D the generated faces are incorrectly facing towards the front while in FIG. 5E the generated face is facing towards the back in alignment with the driving pose in FIG. 5B.

[0070] To qualitatively understand the advantages provided by the methods and systems in accordance with the present embodiments, the performance was evaluated on four quantitative metrics: L1 error, Average Euclidean distance (AED), Average Keypoint Distance (AKD), and Missing Keypoint Rate (MKR). Specifically, the L1 error calculates the mean absolute difference between the generated frames and ground-truth videos. Since ground-truth animation videos are typically unavailable, the L1 error can only be applied to evaluate the self-reconstruction of test videos. The AED evaluates the preservation of identity before and after motion transfer and a person re-identification model was used for extracting identity features and a mean L2 norm of the identity features between the generated and ground-truth frame pairs was calculated. The AKD and MKR indicate how well the pose is transferred in the reconstructed videos. The keypoints of the ground-truth and generated videos are estimated by using a publicly available pose estimation model where the AKD can be obtained by calculating the average distance between the corresponding human body landmarks and the MKR is the missing rate of the landmarks by checking the presence of keypoints for a pair of ground-truth and generated frames.

[0071] The methods and systems in accordance with the present embodiments was quantitatively compared with conventional methods for human image animation using source images and driving sequences from the TED Talks dataset, the Tai Chi HD dataset and the iPER dataset and where the conventional methods included a first and a second first-order motion model identified as FOMM and FOMM-AA. For a fair

comparison, the FOMM and FOMM-AA models were retrained on the three benchmark datasets at 256x256 resolution. All the models were trained under the default hyperparameter settings of the training schedule associated therewith and the performance and video quality were evaluated both on self-reconstruction and cross-video animation.

[0072] The quality of generated videos for self-reconstruction were evaluated with source and driving frames taken from the same video (and hence the same identity). The first frame of a test video was selected as the source frame, while the other frames of the same video were utilized as the driving frames. Using such provides ground-truth for synthesized frames. The quantitative results of the method in accordance with the present embodiments in comparison with state-of-the-art (SOTA) methods are shown in Table 1.

Table 1

	TED Talks (256)				Tai Chi HD (256)				iPER (256)			
	L1↓	AED↓	MKR↓	AKD↓	L1↓	AED↓	MKR↓	AKD↓	L1↓	AED↓	MKR↓	AKD↓
FOMM-AA	0.0264	0.1166	0.0086	2.6573	0.0462	0.1507	0.0263	5.2372	0.0166	0.0567	0.0068	1.5992
FOMM	0.0303	0.1424	0.0102	3.8112	0.0606	0.1667	0.0299	6.5769	0.0226	0.0689	0.0102	2.1754
Present Emb.	0.0316	0.1478	0.0071	2.1255	0.0593	0.1764	0.0137	3.6630	0.0206	0.0828	0.0056	1.1591
Present (128)	0.0317	0.1444	0.0085	1.1372	0.0528	0.1705	0.0132	1.7348	0.0197	0.0826	0.0065	0.5935

[0073] It can be seen from Table 1 that the model in accordance with the present embodiments achieved mixed but generally superior performance on pose transfer, with AKD and MKR values lower than the other methods. However, it had slightly higher values on L1 and AED, which relate to image quality and identity preservation, respectively. There are likely two main reasons for the slightly higher L1 and AED values. The first reason is the resolution for training. Due to the structures of the main components of the system in accordance with the present embodiments having limited GPU memory, 128x128 was used as input while 256x256 was used as output under the

training scheme. This meant that the system in accordance with the present embodiments needed to predict higher-resolution appearance of human body regions from lower-resolution input. The second reason is the unstable DensePose annotation. Since the IUV maps of DensePose are predicted frame by frame, the input IUV map of the model can be unstable across the video sequence. Consequently, the generated frames will also slightly deviate from the ground truth. It is believed that the performance on image quality in terms of L1 and AED can be further improved by addressing these issues in actual implementation.

[0074] To evaluate the performance of the methods and systems in accordance with the present embodiments on cross-video animation, fifty source-driving video pairs were randomly selected from the test set of each benchmark. The source-driving video pairs were then fixed across different methods, including the method in accordance with the present embodiments. The first frame of the source video was regarded as the source image and was animated by a driving video. Table 2 shows quantitative results in terms of AED, AKD and MKR in comparison with state-of-the-art methods. Since a ground-truth for cross-video animation is not available, the L1 error under the animation settings was not able to be evaluated. To assess identity preservation, the re-identification feature of the source frame as ground-truth was repeated n times to obtain the AED (in which n refers to the number of frames in a driving sequence). Moreover, to evaluate the quality of landmark alignment between the animated video and the driving video, the landmarks of the driving frames are predicted and treated as ground truth to calculate AKD and MKR.

Table 2

	TED Talks (256)			Tai Chi HD (256)			iPER (256)		
	AED↓	MKR↓	AKD↓	AED↓	MKR↓	AKD↓	AED↓	MKR↓	AKD↓
FOMM-AA	0.3391	0.0159	7.2375	0.2988	0.0542	12.0688	0.2076	0.0207	5.7536
FOMM	0.2484	0.0496	24.1217	0.2337	0.0776	19.5635	0.1518	0.0193	11.2709
Present Emb.	0.3515	0.0186	3.6755	0.3097	0.0171	3.9812	0.2128	0.0053	1.4138
Present (128)	0.3514	0.0192	1.8588	0.3030	0.0172	1.9308	0.2126	0.0058	0.7212

[0075] The quantitative results in Table 2 show that the model in accordance with the present embodiments can achieve better performance on motion transfer but not as well in identity preservation. The AKD scores of the model in accordance with the present embodiments are lower than for other methods by a large margin across all datasets, with MKR also substantially better for the Tai Chi HD and the iPER datasets. Although the FOMM model achieved a lower AED value, the MKR and AKD values were higher. It was observed that the FOMM model in multiple instances failed to significantly animate the source image, with facial and human body regions remaining mostly static. This lack of motion transfer may explain why the AED value can easily remain low. In comparison to the FOMM-AA model, the model in accordance with the present embodiments achieved competitive AED performance and broadly outperformed the FOMM-AA model in terms of AKD as well as MKR (sans TED Talks).

[0076] Referring to FIGs. 6A, 6B, 6C, 6D, 7A, 7B, 8A and 8B, qualitative animation results from the TED Talks datasets FIGs. (6A to 6D), the Tai Chi HD datasets (FIGs. 7A and 7B), and the iPER datasets (FIGs. 8A and 8B) are shown. It can be seen from these animation results that the methods and systems in accordance with the present embodiments can animate both half and full human body images with better geometry and details (see FIGs. 6A and 6B) and large variations in pose for front-to-back view and self-occlusion (see FIGs. 7A and 7B), while preserving the stability and consistency of animated video sequence (see FIGs. 8A and 8B).

[0077] From the images in FIGs. 6A, 6B, 6C, 6D, 7A, 7B, 8A and 8B it can be seen that the methods and systems in accordance with the present embodiments provides strengths for motion transfer in the three aspects: correct geometry and details is recovered, full-body motion is transferred with correct back-to-front view and self-occlusion, and video animation stability and consistency is preserved.

[0078] For results from the TED Talks dataset, the arm and hand pose of the source person are correctly transferred (see FIGs. 6A and 6B) exemplifying the recovery of correct geometry and details in the images of motion transfer by the methods and systems in accordance with the present embodiments. Due to the semantic ambiguity of the unsupervised regions detected by the FOMM-AA and FOMM models in the same figures, it can be seen that the driving pose cannot be correctly transferred.

[0079] As seen in FIGs. 7A and 7B, the methods and systems in accordance with the present embodiments enable full-body motion while addressing issues of back-to-front view and self-occlusion. As seen in the images of motion transfer in accordance with the present embodiments of FIG. 7A, the model can better handle front-to-back view motion transfer in the Tai Chi HD dataset. In FIG. 7B, it can be seen that self-occlusion can largely degrade the image quality of the FOMM-AA and FOMM models while the motion transfer in accordance with the present embodiments can handle self-occlusion cases well with correct geometry.

[0080] The examples from the iPER dataset in FIGs. 8A and 8B show driving sequences with turning-around and large variation between source and driving images. It can be observed that motion transfer in accordance with the present embodiments generates a more stable video sequence in which the pose is correctly aligned with the driving video as compared to the FOMM and FOMM-AA models, thereby preserving video animation stability and consistency.

[0081] To further validate the effectiveness and contributions of the methods and systems in accordance with the present embodiments, ablation analysis was performed on the model by incrementally introducing key modules or ideas to a baseline. The detailed quantitative results of self-reconstruction and animation is reported on the TED Talks dataset in Table 3 and seen in the image set 900 of FIG. 9. Specifically, to investigate the influence of introducing DensePose annotation, a baseline model (Baseline+Blender 930) was trained by directly using the DensePose IUUV map as input. Under this setting, the RGB source image 910 and source DensePose 912 and driving DensePose 920 with twenty-four human body parts (72 dimensions) were concatenated and directly fed into the generator for pose transfer. For each DensePose IUUV map (source DensePose 912 or driving DensePose 920), twenty-four human body parts are parsed separately. Specifically, the I-component in the IUUV map is encoded to twenty-four semantic labels using one-hot encoding, thereby converting the I-component to twenty-four dimensions. The U and V components are also split into twenty-four channels of feature maps according to the twenty-four human body parts, respectively. Thus, the three-channel IUUV map is encoded to seventy-two dimensional features.

[0082] The BlenderNet has the same structure as the BlenderNet module 340 (FIGs. 3 and 4D) and is adapted to synthesize high-resolution frames. The 2D+Blender 940 represents a model where multi-scale feature warping in accordance with the present embodiments is utilized in both the 2D branch 320 and the BlenderNet module 340 (i.e., the 2D+Blender 940 represents the image generated by processing the source image and the driving image through the MotionNet module 310 and the 2D branch (i.e., the multi-scale feature warping module 320) and the BlenderNet module 340). The 2.5D+Blender 950 represents a model where multi-scale feature warping in accordance with the present embodiments is utilized in both the 2.5D branch of the neural texture mapping module

330 and the BlenderNet module 340 (i.e., the 2.5D+Blender 950 represents the image generated by processing the source image and the driving image through the MotionNet module 310 and the 2.5D branch (i.e., the neural texture mapping module 330) and the BlenderNet module 340). The full model 960 refers to multi-scale feature warping in accordance with the present embodiments being utilized in both the 2D branch 320 and the 2.5D branch 330 as well as in the BlenderNet module 340, which is the model in accordance with the methods and systems of the present embodiments (i.e., the full model 960 represents the image generated by processing the source image and the driving image through the MotionNet module 310, the multi-scale feature warping module 320, the neural texture mapping module 330 and the BlenderNet module 340).

Table 3

	Self-reconstruction				Animation		
	L1↓	AED↓	MKR↓	AKD↓	AED↓	MKR↓	AKD↓
<i>Baseline + BlenderNet</i>	0.0363	0.1727	0.0108	2.1405	0.3611	0.0212	3.1390
<i>2D + BlenderNet</i>	0.0319	0.1486	0.0088	1.9833	0.3481	0.0176	3.3809
<i>2.5D + BlenderNet</i>	0.0349	0.2163	0.0089	3.5312	0.4592	0.0171	5.3624
<i>2D + 2.5D + BlenderNet</i>	0.0316	0.1478	0.0071	2.1255	0.3514	0.0192	1.8588

[0083] It can be seen from Table 3 that the baseline+BlenderNet model 930 achieved good performance for self-reconstruction with lower L1 and AED, but relatively lower performance in animation. It can be regarded as a strong baseline and demonstrates the effectiveness of using DensePose as conditional supervision for motion transfer.

[0084] The 2D+BlenderNet model 940 shows superior performance with lower reconstruction loss and better pose transfer for self-reconstruction in comparison with other baseline models, but a slightly higher value in AKD for animation. This demonstrates that the 2D branch 320 with multi-scale feature warping preserves the appearance by leveraging the predicted dense motion flow. The reason for the higher AKD is related to the quality and stability of DensePose for calculating motion flow.

The translated feature only with feature warping in the 2D branch may not be robust to corrupted DensePose IUV map.

[0085] The 2.5D+BlenderNet model 950 achieved relatively poor performance in animation, with image quality degraded because of wrongly translated pose. It is believed that the reasons for this poor performance include both the single branch training and the model hyperparameter settings. In regards to the single branch training issue, the 2.5D+BlenderNet model 950 with differentiable texture mapping may not benefit from single-stream training to render a high-quality appearance. From observations, the single 2.5D branch 330 may not work well in transferring the source pose to a driving pose, especially for cross-identity pose transfer. In regards to the model hyperparameter setting, in order to implement a fair comparison, we adopted the same network structure for the 2.5D branch only as the structure for the full model. However, for the single 2.5D branch, the hyperparameter settings for modelling the diversity of motion change seem to be insufficient for texture atlas and DensePose translation. The model capability is limited by the length of latent vectors of α and ρ , the layers of translation network, the length of neural texture feature vectors, and similar hyperparameter factors.

[0086] The full model 960 achieved better performance in comparison with the other models, demonstrating that the full model can take advantage of both 2D and 2.5D branches 320, 330 (FIG. 3) by combining the features of flow warping and neural texture mapping to produce high-quality appearance and geometry.

[0087] The images of FIG. 9 also show that the 2D+BlenderNet model 940 can preserve more details of the source person, especially for the face. The baseline+BlenderNet model 930 and the 2.5D+BlenderNet model 950, without feature warping, produced blurry images, with facial, hair and clothing details lost. The full

model 960 in accordance with the present embodiments can generate high-quality images with correct poses by jointly training both the 2D branch 320 and the 2.5D branch 330. The results in the full model 960 show that images generated by the full model are more stable with better appearance and geometry.

[0088] Thus, it can be seen that the present embodiments provide methods and systems in accordance with the present embodiments which provide realistic avatar animation through enhanced motion transfer. The present embodiments provide. It can also be seen from the foregoing that appearance and geometry are critical for high-quality pose-guided human image generation and animation. A unified human motion transfer framework has been presented to combine feature warping and neural texture mapping for better recovering the appearance and the driving pose. The methods and systems in accordance with the present embodiments implicitly decouple the neural texture atlas translation and DensePose translation into a two-branch pipeline for correcting the geometric errors in pseudo-ground-truth DensePose annotation. The experimental quantitative results demonstrate that the model in accordance with present embodiments benefits from both appearance and geometry branches by multi-modal feature fusion and joint training. The qualitative results show that the model in accordance with present embodiments can work well for challenging motion patterns with large variations in appearance and pose, such as turning-around, self-occlusion, and front-to-back view motion transfer, as compared to state-of-the-art models.

[0089] The novel highlights of the methods and systems in accordance with the present embodiments are that an integrated framework is provided with component deep networks that exploit complementary strengths in different approaches to motion transfer which together are packaged as a neural texture rendering pipeline. The use of the 2D flow-based appearance transfer branch 320 allows unoccluded visual features to

be better preserved. The use of the 2.5D geometry branch 330 provides greater ability to reason and cope with occlusions and de-occlusions. In addition, although DensePose is used, our networks are able to predict flows and texture changes in regions outside the DensePose skin silhouettes to cater for hair and clothing.

[0090] Therefore, in accordance with the present embodiments, methods and systems have been provided which provide motion transfer to avatars in a virtual environment which overcome drawbacks of current state of the art methods and systems and provides a more realistic motion of the avatars. As there has been a surge of publicity and growing interest in the Metaverse and other forms of virtual worlds, including 3D virtual gaming environments and existing avatars are not entirely realistic looking and users do not have fine-grained control in the body motion of such avatars, the methods and systems in accordance with the present embodiments invention can be used to create photorealistic avatars with direct transfer of body motion from users. In addition, with the growth of video conferencing expected to continue for ease of rapid visual interactions globally, users may want to deploy a self-avatar or digital twin for convenience that exactly resembles the user, but which is properly groomed and wearing appropriate clothing. This allows the user to remain casually dressed and avoid the effort of grooming, while enabling the user to be suitably dressed while offering the option to choose between multiple self-avatars in different clothing, as appropriate for the occasion.

[0091] While exemplary embodiments have been presented in the foregoing detailed description of the present embodiments, it should be appreciated that a vast number of variations exist. It should further be appreciated that the exemplary embodiments are only examples, and are not intended to limit the scope, applicability, operation, or configuration of the invention in any way. Rather, the foregoing detailed description will provide those skilled in the art with a convenient road map for implementing

exemplary embodiments of the invention, it being understood that various changes may be made in the function and arrangement of steps and method of operation described in the exemplary embodiments without departing from the scope of the invention as set forth in the appended claims.

CLAIMS

What is claimed is:

1. A method for avatar animation in a virtual environment, the method comprising:

receiving first data of at least one source image corresponding to an avatar to be animated in the virtual environment;

receiving second data corresponding to the desired animation; and

combining the first data and the second data to animate the avatar, wherein the step of combining the first data and the second data comprises processing the first and second data to generate a motion-transferred image in response to both latent-feature warping of appearance variation between the first and second data and neural texture mapping of coarse spatial differences between the first and second data.

2. The method in accordance with Claim 1 wherein the latent-feature warping of appearance variation between the first and second data comprises two-dimensional latent-feature warping of appearance variation between the first and second data.

3. The method in accordance with Claim 1 wherein the neural texture mapping of coarse spatial differences between the first and second data comprises occlusion awareness neural texture mapping of coarse spatial differences between the first and second data.

4. The method in accordance with any of Claims 1 to 3 wherein the neural texture mapping of coarse spatial differences between the first and second data

comprises decoupling the appearance variation between the first and second data from the motion variations between the first and second data.

5. The method in accordance with any of Claims 1 to 4 wherein the first data comprises a plurality of pixels forming the at least one source image and the second data comprises driving frames of one or more images forming the desired animation, the method further comprising before the combining step, a step of predicting a dense flow field for regions outside a virtual skin of the avatar to be animated in the virtual environment in response to the plurality of pixels of the at least one source image and the driving frames of the one or more images forming the desired animation.

6. The method in accordance with Claim 5 further comprising before the predicting step, a step of pre-processing the plurality of pixels forming the at least one source image to generate an inferred mapping of a three-dimensional surface-based representation of the avatar to be animated in the virtual environment as a pre-processed source image, and wherein the predicting step comprises predicting the dense flow field for the regions outside the virtual skin of the avatar to be animated in the virtual environment in response to the pre-processed source image, the plurality of pixels of the at least one source image and the driving frames of the one or more images forming the desired animation.

7. The method in accordance with Claim 6 wherein generating the motion-transferred image in response to the neural texture mapping of coarse spatial differences between the first and second data comprises translating one or more texture atlases

generated from mapping the plurality of pixels of the at least one source image to a canonical texture space to predict appearance variations.

8. The method in accordance with Claim 6 or Claim 7 wherein generating the motion-transferred image in response to the neural texture mapping of coarse spatial differences between the first and second data comprises translating the inferred mapping of the three-dimensional surface-based representation of the avatar to be animated in the virtual environment to predict coarse spatial variations with occlusion determination.

9. The method in accordance with any of Claims 5 to 8 wherein generating the motion-transferred image in response to the latent-feature warping of appearance variation between the first and second data comprises generating the motion-transferred image in response to the latent-feature warping of the predicted dense flow field.

10. A system for avatar animation in a virtual environment which receives first data of at least one source image corresponding to an avatar to be animated in the virtual environment and second data corresponding to the desired animation, the system comprising:

a multi-scale feature warping module configured to latent-feature warp appearance variations between the first and second data;

a neural texture mapping module configured to neural texture map coarse spatial differences between the first and second data; and

a BlenderNet module coupled to the multi-scale feature warping module and the neural texture mapping module and configured to generate a motion-transferred image

to animate the avatar in the virtual environment in response to both the latent-feature warp appearance variations between the first and second data and the neural texture map of coarse spatial differences between the first and second data.

11. The system in accordance with Claim 10 wherein the multi-scale feature warping module is configured to two-dimensional latent-feature warp appearance variations between the first and second data.

12. The system in accordance with Claim 10 wherein the neural texture mapping module is configured to occlusion awareness neural texture map the coarse spatial differences between the first and second data.

13. The system in accordance with any of Claims 10 to 12 the neural texture mapping module is configured to decouple the appearance variation between the first and second data from the motion variations between the first and second data.

14. The system in accordance with any of Claims 10 to 13 wherein the first data comprises a plurality of pixels forming the at least one source image and the second data comprises driving frames of one or more images forming the desired animation, the system further comprising a MotionNet module configured to predict a dense flow field for regions outside a virtual skin of the avatar to be animated in the virtual environment in response to the plurality of pixels of the at least one source image and the driving frames of the one or more images forming the desired animation.

15. The system in accordance with Claim 14 further comprising a pre-processing module configured to generate an inferred mapping of a three-dimensional surface-based representation of the avatar to be animated in the virtual environment as a pre-processed source image in response to the plurality of pixels forming the at least one source image, and wherein the MotionNet module is configured to predict the dense flow field for the regions outside the virtual skin of the avatar to be animated in the virtual environment in response to the pre-processed source image, the plurality of pixels of the at least one source image and the driving frames of the one or more images forming the desired animation.

16. The system in accordance with Claim 15 wherein the pre-processing module comprises a DensePose pre-processing module, and wherein the DensePose pre-processing module is configured to generate an inferred mapping of a three-dimensional surface-based representation of the avatar to be animated in the virtual environment as a pre-processed source image utilizing DensePose.

17. The system in accordance with Claim 15 or Claim 16 wherein the neural texture mapping module is configured to generate the motion-transferred image in response to the neural texture mapping of coarse spatial differences between the first and second data by translating one or more texture atlases generated from mapping the plurality of pixels of the at least one source image to a canonical texture space to predict appearance variations.

18. The system in accordance with Claim 16 or Claim 17 wherein the neural texture mapping module is configured to generate the motion-transferred image in

response to the neural texture mapping of coarse spatial differences between the first and second data by translating the inferred mapping of the three-dimensional surface-based representation of the avatar to be animated in the virtual environment to predict coarse spatial variations with occlusion determination.

19. The system in accordance with any of Claims 14 to 18 wherein the BlenderNet module is coupled to the MotionNet module and is configured to generate the motion-transferred image in response to the latent-feature warping of appearance variation between the first and second data by generating the motion-transferred image in response to the latent-feature warping by the multi-scale feature warping module of the predicted dense flow field.

20. A non-transitory computer-readable medium storing software instructions that, when executed by a processor, cause the processor to perform a method in accordance with any of Claims 1 to 9.