PTUS: Photo-Realistic Talking Upper-Body Synthesis via 3D-Aware Motion Decomposition Warping

Luoyang Lin[1], Zutao Jiang[2], Xiaodan Liang[1, 2, 3*], Liqian Ma[4], Michael C. Kampffmeyer[5], Xiaochun Cao[1*]

[1]Shenzhen Campus of Sun Yat-sen University
[2]Mohamed bin Zayed University of Artificial Intelligence
[3]DarkMatter AI Research, Guangzhou, China
[4]ZMO AI Inc. [5]UiT The Arctic University of Norway
{cooluoluoluo, taozujiang, xdliang328}@gmail.com, liqianma.scholar@outlook.com, michael.c.kampffmeyer@uit.no, caoxiaochun@mail.sysu.edu.cn

Abstract

Talking upper-body synthesis is a promising task due to its versatile potential for video creation and consists of animating the body and face from a source image with the motion from a given driving video. However, prior synthesis approaches fall short in addressing this task and have been either limited to animating heads of a target person only, or have animated the upper body but neglected the synthesis of precise facial details. To tackle this task, we propose a **P**hoto-realistic **T**alking **U**pper-body **S**ynthesis method via 3D-aware motion decomposition warping, named **PTUS**, to both precisely synthesize the upper body as well as recover the details of the face such as blinking and lip synchronization. In particular, the motion decomposition mechanism consists of a face-body motion decomposition, which decouples the 3D motion estimation of the face and body, and a local-global motion decomposition, which decomposes the 3D face motion into global and local motions resulting in the transfer of facial expression. The 3D-aware warping module transfers the large-scale and subtle 3D motions to the extracted 3D depth-aware features in a coarse-to-fine manner. Moreover, we present a new dataset, Talking-UB, which includes upper-body images with high-resolution faces, addressing the limitations of prior datasets that either consist of only facial images or upper-body images with blurry faces. Experimental results demonstrate that our proposed method can synthesize high-quality videos that preserve facial details, and achieves superior results compared to state-of-the-art cross-person motion transfer approaches. Code and collected dataset are released in https://github.com/cooluoluo/PTUS.

Introduction

Image animation is a challenging problem in video creation and finds practical applications in settings such as video conferencing, news reports, and role-playing video games. This has given rise to a multitude of synthesis methods that, based on a source image and a driving video, aim to transfer the motion of the person in the driving video to the source image while preserving the appearance information of the person in the source image (Wang, Mallya, and Liu 2021; Siarohin et al. 2021; Hong et al. 2022). These synthesis approaches either focus only on facial animations (Wang, Mallya, and Liu 2021; Hong et al. 2022), known as talking head synthesis approaches (Fig. 2a), or on body motions without considering facial details (Siarohin et al. 2019b), known as full-body animation approaches (Fig. 2b). This limits their applicability in practical applications, such as news reporting, in which both body motion and facial details matter, giving rise to the challenging talking upper-body synthesis task (Fig. 2c). This task extends the talking head synthesis task to include upper-body motion, as well as facial expressions.

To tackle the talking head problem, warping-based algorithms (Kewei et al. 2022; Hong et al. 2022; Wang, Mallya, and Liu 2021; Doukas, Zafeiriou, and Sharmanska 2021) that learn a dense motion field to warp the source image features, have established themselves as the dominant approach. These methods have recently been extended to lever-

Figure 1: Talking upper-body synthesis results produced by our method trained on the Talking-UB dataset. Our method can transfer both large-scale motions, such as body movement, as well as subtle facial expressions, such as eye blinking and mouth motion, from the driving video to the source image.

*Corresponding author

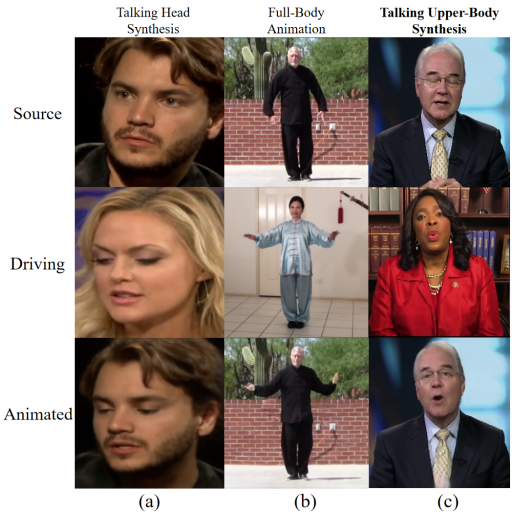

Figure 2: We introduce a new challenging task termed talking upper-body synthesis. (a) Talking head synthesis focuses on facial animation only (Wang, Mallya, and Liu 2021; Hong et al. 2022). (b) Full-body animation ignores facial details (Siarohin et al. 2019b). (c) Talking upper-body synthesis instead can both transfer the motions of the upper-body and face and preserves facial details.

age 3D geometric information in order to improve the synthesis quality of, among others, face rotation (Hong et al. 2022; Wang, Mallya, and Liu 2021). This is done by learning a 3D depth feature map and a depth map in a self-supervised manner. However, these methods generate facial details without considering the motions of the body.

Human body animation, on the other hand, has recently gained attention through the impressive performance of motion transfer approaches (Siarohin et al. 2019a,b, 2021; Wang et al. 2022). While these approaches transfer motions among people, these approaches focus on body movement and are unable to accurately capture the facial details. In summary, existing works either only focus on facial animation, ignoring the body motion, or only on body animation while ignoring the synthesis of facial details.

In this work, we propose a novel **P**hoto-realistic **T**alking **U**pper-body **S**ynthesis method via 3D-aware motion decomposition warping, named **PTUS**, which can animate both face and body with a 3D-aware motion decomposition warping mechanism. The 3D-aware warping module manipulates the 3D features of the source image conditioned on the 3D motions learned from the driving image. To precisely synthesize the upper-body and facial details, we first decompose the 3D motion into the face and body motions with a face-body motion decomposition. We then decompose the face motion into global and local motions with a local-global motion decomposition to facilitate fine-grained warping such as blinking of the eyes via the local face motion. As illustrated in Fig. 1, PTUS can achieve realistic and detailed results for body motion transfer such as body movements (the first column), as well as facial motion transfer such as eyes blinking (the second column) and face rotation (the third column), via the proposed 3D-aware motion decomposition warping mechanism. Finally, we address the other roadblock that currently exists for Talking Upper-Body Synthesis, which is the lack of high-resolution faces in existing datasets. To address this issue, we collect a new dataset, named Talking-UB, which contains upper-body images with high-resolution faces.

The main contributions are three-fold:

• We propose the first method for synthesizing a talking upper-body, **PTUS**, which animates the body and face from a source image with the motion from a given driving video.

• A 3D-aware motion decomposition warping mechanism is proposed to precisely synthesize facial as well as upper-body details. The 3D-aware warping transfers 3D motion from the driving image to the source image in a coarse-to-fine manner. The face-body motion decomposition mechanism estimates the 3D motions of the face and body separately, while the local-global motion decomposition mechanism decomposes the 3D face motion into global and local motions. Benefiting from the 3D-aware motion decomposition warping mechanism, our model can both learn large-scale motions such as body movement and subtle motions such as blinking eyes.

• We create a more challenging benchmark for body and face animation, named Talking-UB, containing upper-body images with high-resolution faces.

Related Work

Image-to-Image Translation. The goal of image-to-image translation is to map an existing image to a new one with a specific style or characteristic, while maintaining the content. With the advent of generative models (Goodfellow et al. 2014; Kingma and Welling 2013), this topic has been extensively studied. To translate image style, early works (Isola et al. 2017; Wang et al. 2018) propose to learn a deterministic one-to-one mapping by leveraging a U-Net trained in an adversarial way. (Zhu et al. 2017; Huang et al. 2018; Lee et al. 2018; Ma et al. 2019) further extend the task to the multimodal setting where one image can be translated to different styles via one-to-many mappings. Recent works further explore instance-level translation (Mo, Cho, and Shin 2019), contrastive learning (Park et al. 2020), transformer architecture (Kim et al. 2022), *etc.* However, image-to-image methods focus primarily on appearance transfer instead of motion transfer, which is crucial in our talking upper-body synthesis task.

Motion Transfer. Motion transfer aims to transfer the motion of one person to another. Early methods designed for this task make explicit use of motion guidance. For example, (Ma et al. 2017; Zhu et al. 2019; Ren et al. 2020; Zhang et al. 2020; Zhou et al. 2021) use OpenPose (Cao et al. 2017) to extract human body keypoints as guidance to transfer body pose. However, such domain specific guidance requires prior knowledge of animated objects and extra annotations. Instead, recent works have attempted to learn the motion representation of an image directly by disentangling identity and pose in an unsupervised manner. (Siarohin et al. 2019a) predict keypoints from source and driving images and employ the keypoints to learn a dense motion field for warp-

ing source features in a self-supervised manner. (Siarohin et al. 2019b) adopt a first-order Taylor expansion to improve the flow between the source features and driving features. (Siarohin et al. 2021) propose a novel motion representation based on a heatmap to improve the quality of motion transfer. (Zhao and Zhang 2022) propose a thin-plate spline motion prediction and estimate a multi-resolution occlusion mask to overcome large pose differences between the source and driving images. Unlike (Siarohin et al. 2019b, 2021; Zhao and Zhang 2022), (Wang et al. 2022) manipulates motion transfer in latent space rather than explicit representations such as keypoints or regions. Although these motion transfer methods can be applied to the proposed talking upper-body synthesis task, they focus on body motion while neglecting facial expression motion, leading to sub-optimal performance.

Video-driven Talking Head Synthesis. The task of video-driven talking head generation aims to animating the face from a source image with the motion from a driving video. Burkov et al. proposes a head animation framework by encoding pose expressions into a latent space (Burkov et al. 2020). (Kewei et al. 2022) adopts an embedding network and a driving network to learn an embedded face and the warping between the embedded face and the target image. (Hong et al. 2022) leverages a depth map to improve keypoint estimation and a cross-modal attention mechanism to boost performance for motion transfer. (Wang, Mallya, and Liu 2021) warps 3D features from source image space to driving image space, allowing rotating and translating operations. (Yao et al. 2020) introduces a facial animation using a 3DMM parametric model (Blanz and Vetter 1999) as guidance. (Zhang et al. 2019) models Face Reenactment by disentangling appearance and shape information in latent spaces. (Zakharov et al. 2019) utilizes a encoder to extract appearance code and landmarks as pose-guided information. (Zhao, Wu, and Guo 2021) guides the prediction of the dense motion map using landmarks in a global-local manner. While (Deng et al. 2020) models face animation by disentangling face attributes using StyleGAN (Karras, Laine, and Aila 2019). These talking head methods can synthesize detailed facial expressions, *e.g.*, eye blinking and mouth motion, however, they do not consider body motion which is one key factor in the upper-body scenario. Compared with existing talking head synthesis methods, the proposed talking upper-body synthesis method facilitates face and body animation in a unified framework via a novel 3D-aware motion decomposition.

Method

Overview

In this section, we introduce a novel framework (see Fig. 3) for talking upper-body synthesis that can learn both face and upper-body motion transfer through our proposed 3D-aware motion decomposition. Given a source image S and a driving video $D = \{D_1, D_2, ..., D_n\}$ consisting of n frames, the proposed method aims to transfer the motion of the person in the driving video to the person in the source image and generates a target video $T = \{T_1, T_2, ..., T_n\}$. The identity in the target video T is the same as in the source image but inherits the motion of the person from the driving video D.

As shown in Fig. 3 (b), the proposed motion decomposition mechanism consists of a face-body motion decomposition module and a local-global motion decomposition module. The face-body decomposition module first decouples the 3D motion estimation of the face and upper-body. To learn subtle motions such as eye blinking, the local-global motion decomposition module decomposes the predicted 3D face motion into a local face motion m_l^{fa} and a global face motion m_g^{fa}. The 3D-aware warping module first transfers the learned large-scale 3D motions of the face and upper-body to the extracted source 3D features of the source image and then transfers the decomposed 3D face motion to the initial warped features for the fine-grained warping. After obtaining the refined 3D facial features, we combine them with the body features and a generation module is utilized to decode the combined features to the target image.

We will first introduce our proposed face-body motion decomposition and the local-global motion decomposition. The 3D-aware warping module is then presented before we finally provide the loss functions for training the framework.

Face-Body Motion Decomposition

Unlike previous talking head synthesis and motion transfer approaches, which estimate the motion of the person as a whole, the proposed face-body motion decomposition module estimates the 3D motion of the face and upper-body separately. The key insight behind the face-body motion decomposition is that the movements of different parts in the articulated object (*e.g.*, the face and upper-body of a person) are highly independent. Modeling the movement of different parts with one motion representation is therefore inappropriate. In the following, we describe how our framework predicts the 3D source and driving keypoints and leverages them to estimate the 3D motion of the face and upper-body.

3D Keypoints Prediction. Taking a source image S and its face S^{fa} as input, the keypoint detector (Wang, Mallya, and Liu 2021) is adopted to extract the keypoints of the source image in the canonical space to represent a person's geometric signature in a neutral pose and expression. To transform 3D keypoints from the canonical space to the observation space, we apply pose estimators to encode pose information and minor motion modules to represent a person's facial expression $(face)$ or small movement $(body)$.

We predict two groups of keypoints from the source image and its face crop to represent body and face structures, one for the upper-body (including the head) and the other for the face, which allows us to estimate facial and upper-body motions separately. Specifically, the face keypoints x^{fa} and upper-body keypoints x^{bo} in the canonical space can be predicted as follows:

$$x^{fa} = F_{kp}^{fa}(S^{fa}), \quad x^{bo} = F_{kp}^{bo}(S), \tag{1}$$

where F_{kp}^{fa} and F_{kp}^{bo} are two keypoint detectors with the same network architectures for the face and upper-body, respectively.

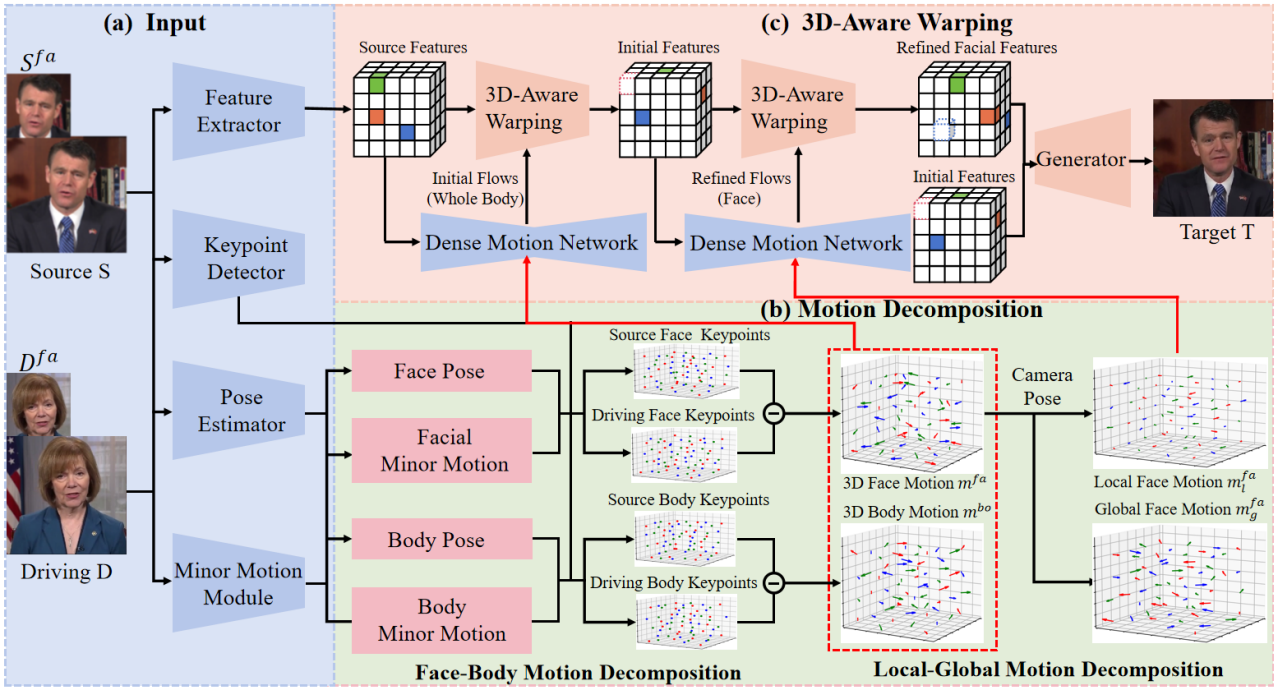

Figure 3: The framework of our PTUS model for talking upper-body synthesis. (a) Inputs: Source and driving images and corresponding face crops. (b) The motion decomposition module consists of the face-body motion decomposition and the local-global motion decomposition. The face-body motion decomposition mechanism estimates the 3D motions for face and body separately, while the local-global motion decomposition mechanism decomposes the 3D face motion into a local face motion and global face motion. The local face motion is responsible for fine-grained warping such as eyes blinking. (c) The 3D-aware warping module first transfers 3D motions from the driving image to the source image and then generates the target image with the initial and refined facial features.

To transform 3D keypoints from the canonical space to the observation space, we adopt two pose estimators (Wang, Mallya, and Liu 2021) with the same network architecture to estimate the camera poses of the face and upper-body in the source and driving images, respectively. The transformed source keypoints in observation space of the source image can be written as:

$$x_S^{fa} = R_S^{fa} x^{fa} + t_S^{fa}, \quad x_S^{bo} = R_S^{bo} x^{bo} + t_S^{bo}, \tag{2}$$

where R_S and t_S denote the rotation matrix and the translation vector predicted with the pose estimators from the source image, respectively. The superscript indicates if the face (fa) or upper-body (bo) pose estimator is leveraged. To transfer the 3D motions from the driving images to the source image, we also transform the keypoints extracted from the source image to the observation space of the driving images. The transformed driving keypoints in observation space of the driving images can be written as:

$$x_D^{fa} = R_D^{fa} x^{fa} + t_D^{fa}, \quad x_D^{bo} = R_D^{bo} x^{bo} + t_D^{bo}, \tag{3}$$

where R_D and t_D denote the predicted rotation matrix and the translation vector from the driving images, respectively.

Unfortunately, the pose estimator can only handle the large-scale 3D motions such as the face and body rotations, not the subtle expression motions or small movement such as the blinking of the eye. As a remedy, we adopt two minor motion modules (Wang, Mallya, and Liu 2021) with the same network architecture to learn facial expression and small movements and apply them to the keypoints in observation space. Formally,

$$\begin{aligned} x_S^{fa} &= R_S^{fa} x^{fa} + t_S^{fa} + e_S^{fa}, \\ x_S^{bo} &= R_S^{bo} x^{bo} + t_S^{bo} + e_S^{bo}, \\ x_D^{fa} &= R_D^{fa} x^{fa} + t_D^{fa} + e_D^{fa}, \\ x_D^{bo} &= R_D^{bo} x^{bo} + t_D^{bo} + e_D^{bo}, \end{aligned} \tag{4}$$

where e_S and e_D represent the facial expression deformation (fa) or small body movements (bo) obtained from the source and driving images, respectively. After obtaining the source and driving keypoints, we calculate the 3D motions of the face and upper-body as follows:

$$m^{fa} = x_D^{fa} - x_S^{fa}, \quad m^{bo} = x_D^{bo} - x_S^{bo}, \tag{5}$$

where m^{fa} and m^{bo} are the 3D motions of the face and upper-body, respectively. Inspired by FOMM (Siarohin et al. 2019b), we learn the keypoints to represent 3D motions following a self-supervised strategy.

Local-Global Motion Decomposition

Facial expressions play an important role in the talking upper-body task. In contrast to 3D motions of the face, such

as face rotation, facial expression deformations such as eye blinking are subtle. Consequently, training the talking upper-body synthesis model directly will lead to the optimization of the pose estimator, while ignoring the transfer of facial expressions. To alleviate this issue, we propose a local-global motion decomposition mechanism to decompose the 3D face motion into global and local motions. The decomposed local motions can be utilized to generate refined flows for fine-grained 3D-aware warping.

Specifically, we decompose the 3D face motion m^{fa} into global and local components as follows:

$$m^{fa} = m_g^{fa} + m_l^{fa}, \tag{6}$$

where m_g^{fa} and m_l^{fa} denote the 3D global and local motion of the face. We calculate the local face motion with a keypoint alignment technique as follows:

$$m_l^{fa} = e_D^{fa} - R_D^{fa}(R_S^{fa})^{-1} e_S^{fa}. \tag{7}$$

As shown in Eq. 7, we project the expression representation from the source image pose to the driving image pose and this helps our model learn facial expression independent of the pose change. Benefiting from the decomposed local and global motion, our model can transfer the subtle expression deformation from the driving image to the source image.

3D-Aware Warping

Unlike previous methods (Wang, Mallya, and Liu 2021; Siarohin et al. 2019b), which warp features focusing on the principal component of the motions and thus struggle to perform fine-grained warping, our 3D-aware warping module warps the depth-aware 3D features extracted from the source image using a feature extractor (Wang, Mallya, and Liu 2021) in a coarse-to-fine manner. It first transfers the large-scale 3D motions of the whole body to the extracted 3D source features to obtain the initial warped features, and then transfers the subtle 3D local motions of the face to refine the initial warped features for fine-grained 3D-aware warping. Specifically, a dense motion network (Wang, Mallya, and Liu 2021) is first adopted to learn initial dense optical flows w^{init} by taking the body motions as input. Then, we apply w^{init} to the 3D canonical coordinates C_{3D}^{std} of the 3D source features f_s and obtain the 3D initial deformed coordinates C_{3D}^{init}. The 3D initial warped features f^{init} can be obtained as follows:

$$f^{init} = F_g(f_s, C_{3D}^{init}). \tag{8}$$

where F_g denotes the 3D-Aware warping operation, which is implemented using the $Grid\_Sample$ algorithm (Jaderberg et al. 2015). Since the initial warped 3D features are associated with the global motion like body movement or head tilt, ignoring minor facial motion like eyes blinking, we decompose the face motion into a global and a local motion and utilize the local motion to refine the initial feature f^{init}. We adopt a dense motion network (Wang, Mallya, and Liu 2021) to infer a refined dense optical flow w^{re} by taking the local face motion as input. We apply w^{re} to the 3D canonical coordinates C_{3D}^{std} and obtain the 3D refined deformed coordinates C_{3D}^{re}. Finally, the 3D refined warped features are computed as follows:

$$f^{re} = F_g(f^{init}, C_{3D}^{re}) \,. \tag{9}$$

Image Generator. Once the refined 3D warped features are obtained, the image generator (Wang, Mallya, and Liu 2021) first projects them back to 2D. The projected 2D image features are then multiplied by the occlusion mask indicated from the 3D source features and keypoints in the dense motion network. Finally, the masked 2D image features are decoded to obtain the target image.

Inference. In practice, we observe that some parts of the upper-body (e.g., clothing) can blend into the background leading to sub-par generation results for a few instances. To address this issue, we initially utilize a mask generated by MODNet (Ke et al. 2022) to segment the upper-body from the background. We then employ Magic Studio for background inpainting and seamlessly integrate the upper-body with the inpainted background using the predicted mask.

Loss Function

In the training stage, we select source and driving images with the same identity from the same video to train our model in a self-supervised manner. We leverage the loss function from OSFV, which can be summarized as follows:[1]

$$\begin{aligned} L = \lambda_{rec} L_{rec} + \lambda_G L_G + \lambda_P L_P + \lambda_{cam} L_{cam} \\ + \lambda_E L_E + \lambda_\delta L_\delta + \lambda_{kp} L_{kp} + \lambda_B L_B. \end{aligned} \tag{10}$$

Experiments

Experimental Setup

Dataset. Existing datasets for talking-head synthesis almost exclusively include faces (Chung, Nagrani, and Zisserman 2018; Nagrani, Chung, and Zisserman 2017) and are lacking the upper-body view. Meanwhile, the recently released dataset for upper-body animation, TED-Talk(Siarohin et al. 2021), only includes blurry faces. To fill this gap, we introduce a new dataset, Talking-UB, that contains both high-quality faces as well as the upper-body view and can be leveraged for the talking upper-body synthesis task. More than 200 videos with durations ranging from 1 to 15 minutes are collected from YouTube in the following three resolutions: 1080P, 720P and 360P, covering diverse lightning conditions, clothing, and background. We first select the frames containing persons for each video, ensuring an almost constant background for each motion sequence. To crop the upper-body and faces from each frame, we adopt the human body parsing algorithm Graphonomy (Gong et al. 2019). We resize all cropped frames to 256 $\times$ 256 resolution. Unlike the TED-Talk dataset, the height proportion of the face to the upper-body in the images is between 1/3 and 2/3, resulting in high-quality talking upper-body synthesis data. We divided the dataset into a train set and a test set consisting of 180 videos and 33 videos, respectively. For training, we extract three images per second. For testing, we select the first 100 frames or all frames if there are fewer than 100 frames in the video. This results in 82470 frames for the train set

[1]The full loss details are provided in the supplementary.

Figure 4: Qualitative comparison results of different methods for the same-identity motion transfer.

and 3217 images for the test set. More details about the collected Talking-UB dataset are provided in the supplementary material.

Metric. We evaluate the talking upper-body synthesis results using L_1 distance, average keypoint distance (AKD), average euclidean distance (AED), and face landmark distance (FLD) for the same-identity motion transfer. The L_1 distance is leveraged to calculate the pixel-level distance between generated and ground-truth images. AKD (Cao et al. 2017) is adopted to estimate the distance of the keypoint positions between the generated and ground-truth images. AED (Deng et al. 2019) measures the ability to preserve the identity. FLD (Bulat and Tzimiropoulos 2017) measures the consistency of facial expressions by extracting the landmarks from the generated and ground-truth images. For the cross-identity motion transfer, we adopt the video FID (Wang et al. 2022) to calculate the distance between the generated and real videos.

More experimental details are included in the supplementary material.

Comparison with State-of-the-Art Methods

We compare the proposed PTUS with the two state-of-the-art methods LIA (Wang et al. 2022) and TPSMM (Zhao and Zhang 2022), as well as other baseline methods: FOMM (Siarohin et al. 2019a), MRAA (Siarohin et al. 2021). We evaluate the performance both for the same-identity motion transfer and the cross-identity motion transfer on the Talking-UB dataset[2].

Same-identity Motion Transfer. Quantitative comparison results of FOMM, MRAA, LIA, TPSMM, and PTUS are reported in Table 1. It can be observed that PTUS achieves the best results on the Talking-UB dataset in terms of L_1 distance and face landmark distance. This suggests that PTUS can recover facial details more accurately. Fig. 4 shows the qualitative comparison to FOMM, MRAA, LIA, and TPSMM. As displayed in Fig. 4, PTUS can more accurately capture facial expression motion like eye blinking (Row 1) and body motion (Row 3). Further, PTUS can effectively mitigate the issue of the upper body adhering to the background (Row 2), while FOMM, MRAA and TPSMM cannot.

[2]Additional qualitative results can be found in the supplementary

Methods	FOMM	MRAA	LIA	TPSMM	PTUS
L1 ↓	0.040	0.036	0.141	0.036	**0.035**
AKD ↓	2.87	2.70	9.79	2.46	**2.46**
AED ↓	0.048	0.045	0.085	0.043	**0.043**
FLD ↓	1.652	1.542	3.804	1.377	**1.322**
FID ↓	44.50	39.69	82.63	34.28	**26.61**

Table 1: Quantitative results of different methods for motion transfer.

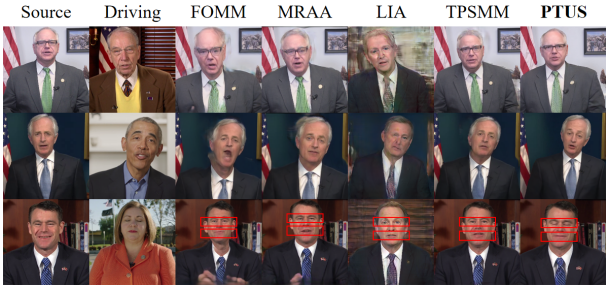

Figure 5: Qualitative comparison of different methods for the cross-identity motion transfer.

Cross-identity Motion Transfer. The task of cross-identity motion transfer is to transfer the motion of one person in the driving images to another person in the source image. The identities in the driving and source images are different. We compare the performance of FOMM, MRAA, LIA, TPSMM, and PTUS for the cross-identity motion transfer on the Talking-UB dataset. To compute the FID score, we randomly select 16 images containing different persons in the test set for each video and generate 528 videos for evaluation. Results are recorded in Table 1. As shown in Table 1, compared with FOMM, MRAA, LIA, and TPSMM, PTUS achieves the best results. This indicates that PTUS can generate more realistic videos. As displayed in Fig. 5, PTUS also achieves better qualitative results compared to other methods. Our model learns more accurate motion transfer for the face and generated facial details like the shape of the mouth and the blinking of the eyes across different identities (Row 3). Further, also in the cross-identity setting, PTUS can effectively mitigate the issue of the upper body adhering to the background (Row 2), compared to FOMM, MRAA and TPSMM. Finally, we observe that PTUS can preserve the identity more faithfully than LIA.

Human Study. We also conduct a human study to evaluate the quality of generated videos for cross-identity motion transfer. Four metrics are considered: **identity** (the consistency between the person of the source image and generated videos), **authenticity** (the authenticity of the generated videos), **facial motion consistency(FMC)** (the consistency between the facial motion of the driving and generated videos), and **body motion consistency(BMC)** (the consistency between the body motions of the driving and generated videos). We randomly select 10 source images and driving videos for cross-identity motion transfer. Results from FOMM, MRAA, LIA, TPSMM, and our method are shown

Methods	Identity	FMC	BMC	Authenticity
FOMM	4.00%	15.75%	16.50%	6.25%
MRAA	10.25%	12.00%	14.25%	6.75%
LIA	1.75%	10.25%	9.75%	13.25%
TPSMM	20.00%	13.75%	12.25%	23.25%
PTUS	**64.00%**	**48.25%**	**47.25%**	**50.50%**

Table 2: Human Study for the cross-identity motion transfer. 40 human raters are asked to evaluate the quality of generated videos. Numbers denote the proportion(%) of users' preferences.

Figure 6: Qualitative comparison of OSFV(Backbone).

in random order and the human raters are asked to select the best video from the perspective of identity, authenticity, facial motion consistency, and body motion consistency. We collect 40 human evaluation results in total and use the proportion (%) of users' preferences as the final scores. Table 2 displays the results of the human study for the cross-identity motion transfer and demonstrates that our method outperforms the state-of-the-art methods in terms of identity, authenticity, facial motion consistency, and body motion consistency.

Methods	L1 ↓	AKD ↓	AED ↓	FLD ↓	FID ↓
OSFV	0.052	4.22	0.050	1.941	27.27
PTUS	**0.035**	**2.46**	**0.043**	**1.322**	**26.61**

Table 3: Compared with OSFV(backbone)

Ablation Study

Compared with OSFV (Backbone). We compare PTUS with the backbone OSFV and study the effectiveness of the proposed motion decomposition for motion transfer on the Talking-UB dataset. Quantitative comparison results are shown in Tab. 3. As shown in Tab. 3, PTUS achieves the best results in terms of L_1 distance, average keypoint distance, average euclidean distance, face landmark distance and FID score. This suggests that PTUS can animate the upper-body in a more precise pose and recover facial details more accurately, maintaining identity authenticity with greater fidelity and generating images with a greater level of realism in motion transfer. The visual results shown in Fig. 6 indicate that our method can synthesis facial motion like face rotation(the second row) and eyes blinking(the third row) more precisely and achieve a higher level of accuracy in capturing body pose(the first row).

Methods	L1 ↓	AKD ↓	AED ↓	FLD ↓	FID ↓
w/o FBLG	0.068	9.85	0.065	3.526	86.44
w/o LGMD	0.067	10.12	0.066	3.631	87.26
PTUS	**0.035**	**2.46**	**0.043**	**1.322**	**26.61**

Table 4: Ablation studies of the FBMD (face-body motion decomposition) and the LGMD (local-global motion decomposition).

Ablation on Motion Decomposition. To further demonstrate the benefit of the proposed face-body motion decomposition (FBMD) and the local-global motion decomposition (LGMD), we compare PTUS to two ablation models where FBMD and LFMD are removed. Results are provided Tab. 4. When removing FBLG (w/o FBLG), the ablation model uses only one pose estimator for the whole upper-body, illustrating the benefit of the proposed decomposition in PTUS that separately models face and body motions. Similarly, the drop in performance when removing LGMD (w/o LGMD), illustrates that the decomposition of facial motions into global and local components in PTUS is required to enhance body and facial motions.

Conclusion

In this paper, we introduce a new challenging problem, namely, talking upper-body animation. As opposed to previous research, talking upper-body animation aims to synthesize *both* the upper body and recover facial details. To tackle this task, we propose a 3D-aware motion decomposition framework for talking upper-body animation. We first decompose the 3D motion estimation of the face and body. To recover the facial expression, we further decompose 3D face motions into global and local motions with a local-global motion decomposition mechanism for fine-grained warping. Finally, as prior datasets either consist of only facial images or upper-body images with blurry images, we propose a new dataset, Talking-UB, which includes upper-body images with high-quality faces. Experimental results demonstrate that our method can produce more realistic and detailed results for the talking upper-body task. [3]

[3]Limitations and future work are included in the supplementary material.

Ethics Statement

Although realistic upper-body synthesis can be useful in applications such as video conferencing, news reporting, and role-playing video games, as with any animation technique, these models can be misused. For example, to generate deep fakes. However, forensic analysis and other manipulation detection methods could mitigate such negative effects.

Acknowledgments

This work was supported in part by National Key R&D Program of China under Grant No. 2020AAA0109700, Guangdong Outstanding Youth Fund (Grant No. 2021B1515020061), National Natural Science Foundation of China (NSFC)under Grant No.61976233, No. 92270122 and No.62025604, Mobility Grant Award under Grant No. M-0461, Shenzhen Science and Technology Program (Grant No. RCYX20200714114642083), Shenzhen Science and Technology Program (Grant No. GJHZ20220913142600001), Nansha Key RD Program under Grant No.2022ZD014 and Sun Yat-sen University under Grant No. 22lgqb38 and 76160-12220011, as well as computation resources and contributions from the industry partner(s).

References

Blanz, V.; and Vetter, T. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*.

Bulat, A.; and Tzimiropoulos, G. 2017. How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks). In *International Conference on Computer Vision*.

Burkov, E.; Pasechnik, I.; Grigorev, A.; and Lempitsky, V. 2020. Neural head reenactment with latent pose descriptors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Chung, J. S.; Nagrani, A.; and Zisserman, A. 2018. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*.

Deng, J.; Guo, J.; Niannan, X.; and Zafeiriou, S. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *CVPR*.

Deng, Y.; Yang, J.; Chen, D.; Wen, F.; and Tong, X. 2020. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Doukas, M. C.; Zafeiriou, S.; and Sharmanska, V. 2021. Headgan: One-shot neural head synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Gong, K.; Gao, Y.; Liang, X.; Shen, X.; Wang, M.; and Lin, L. 2019. Graphonomy: Universal Human Parsing via Graph Transfer Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; and Bengio, Y. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*.

Hong, F.-T.; Zhang, L.; Shen, L.; and Xu, D. 2022. Depth-Aware Generative Adversarial Network for Talking Head Video Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Huang, X.; Liu, M.-Y.; Belongie, S.; and Kautz, J. 2018. Multimodal Unsupervised Image-to-image Translation. In *European Conference on Computer Vision*.

Isola, P.; Zhu, J.; Zhou, T.; and Efros, A. A. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; et al. 2015. Spatial transformer networks. *Advances in Neural Information Processing Systems*.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Ke, Z.; Sun, J.; Li, K.; Yan, Q.; and Lau, R. W. 2022. MODNet: Real-Time Trimap-Free Portrait Matting via Objective Decomposition. In *AAAI*.

Kewei, Y.; Kang, C.; Daoliang, G.; Song-Hai, Z.; Yuan-Chen, G.; and Weidong, Z. 2022. Face2Faceρ: Real-Time High-Resolution One-Shot Face Reenactment. In *European Conference on Computer Vision*. Springer.

Kim, S.; Baek, J.; Park, J.; Kim, G.; and Kim, S. 2022. InstaFormer: Instance-Aware Image-to-Image Translation with Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Lee, H.-Y.; Tseng, H.-Y.; Huang, J.-B.; Singh, M.; and Yang, M.-H. 2018. Diverse Image-to-Image Translation via Disentangled Representations. In *European Conference on Computer Vision*.

Ma, L.; Jia, X.; Georgoulis, S.; Tuytelaars, T.; and Van Gool, L. 2019. Exemplar Guided Unsupervised Image-to-Image Translation with Semantic Consistency. In *International Conference on Learning Representations*.

Ma, L.; Jia, X.; Sun, Q.; Schiele, B.; Tuytelaars, T.; and Van Gool, L. 2017. Pose guided person image generation. In *Advances in Neural Information Processing Systems*.

Mo, S.; Cho, M.; and Shin, J. 2019. InstaGAN: Instance-aware Image-to-Image Translation. In *International Conference on Learning Representations*.

Nagrani, A.; Chung, J. S.; and Zisserman, A. 2017. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*.

Park, T.; Efros, A. A.; Zhang, R.; and Zhu, J.-Y. 2020. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*. Springer.

Ren, Y.; Yu, X.; Chen, J.; Li, T. H.; and Li, G. 2020. Deep image spatial transformation for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019a. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019b. First order motion model for image animation. *Advances in Neural Information Processing Systems*.

Siarohin, A.; Woodford, O. J.; Ren, J.; Chai, M.; and Tulyakov, S. 2021. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Wang, T.-C.; Mallya, A.; and Liu, M.-Y. 2021. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Wang, Y.; Yang, D.; Bremond, F.; and Dantcheva, A. 2022. Latent Image Animator: Learning to Animate Images via Latent Space Navigation. *arXiv preprint arXiv:2203.09043*.

Yao, G.; Yuan, Y.; Shao, T.; and Zhou, K. 2020. Mesh guided one-shot face reenactment using graph convolutional networks. In *Proceedings of the 28th ACM International Conference on Multimedia*.

Zakharov, E.; Shysheya, A.; Burkov, E.; and Lempitsky, V. 2019. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Zhang, P.; Zhang, B.; Chen, D.; Yuan, L.; and Wen, F. 2020. Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhang, Y.; Zhang, S.; He, Y.; Li, C.; Loy, C. C.; and Liu, Z. 2019. One-shot face reenactment. *arXiv preprint arXiv:1908.03251*.

Zhao, J.; and Zhang, H. 2022. Thin-Plate Spline Motion Model for Image Animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhao, R.; Wu, T.; and Guo, G. 2021. Sparse to dense motion transfer for face image animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Zhou, X.; Zhang, B.; Zhang, T.; Zhang, P.; Bao, J.; Chen, D.; Zhang, Z.; and Wen, F. 2021. Cocosnet v2: Full-resolution correspondence learning for image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhu, J.-Y.; Zhang, R.; Pathak, D.; Darrell, T.; Efros, A. A.; Wang, O.; and Shechtman, E. 2017. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*.

Zhu, Z.; Huang, T.; Shi, B.; Yu, M.; Wang, B.; and Bai, X. 2019. Progressive pose attention transfer for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.