Human Image Animation via Semantic Guidance

ABSTRACT

Image animation creates visually compelling effects by animating still source images according to driving videos. Recent work performs animation on arbitrary objects using unsupervised methods and can relatively robustly perform motion transfer on human bodies. However, the complex representation of motion and unknown correspondence between human bodies often lead to issues such as distorted limbs and missing semantics, which make human animation challenging. In this paper, we propose a semantically guided, unsupervised method of motion transfer, which uses semantic information to model motion and identity. Specifically, we use a pre-trained human parsing network to encode the rich and diverse foreground semantic information, thus generating fine details. Secondly, we use a cross-modal attention layer to learn the semantic region's correspondence between human bodies to guide the network in selecting appropriate input features, prompting the network to generate accurate results. Experiments demonstrate that our method outperforms state-of-the-art methods in motion-related metrics, while effectively addressing the problems of semantic missing and unclear limb structures prevalent in human motion transfer. These improvements can facilitate its applications in various fields, such as education and entertainment.

CCS CONCEPTS

• Insert CCS text here • Insert CCS text here • Insert CCS text here

KEYWORDS

Image Animation, Motion Transfer, Semantic Guided, Cross-modal Attention

1 INTRODUCTION

The image animation task sets animation for still images by extracting motion information from video sequences. It has diverse application scenarios, such as video conferencing [1–3], e-commerce [4], and animation production [5]. Given a source image and a driving video, image animation aims to generate a new video that preserves the source image identity information while exhibiting the driving video's motion patterns. The identity here refers to the object's appearance in the source image, and the motion pattern is a series of continuously changing poses [6].

*Article Title Footnote needs to be captured as Title Note

†Author Footnote to be captured as Author Note

WOODSTOCK'18, June, 2018, El Paso, Texas USA

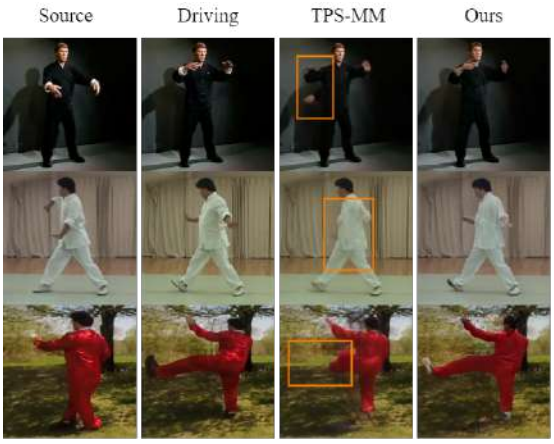

Figure 1. Failure cases from the TPS-MM [13] method. Complex and large motions cause part of the human body to be missing or distorted limbs. Our method improved the issue.

Human image animation can also be understood as the motion transfer of the human body. In recent years, several successful approaches have been proposed for motion transfer tasks [5,7–13]. The latest unsupervised motion transfer methods reconstruct the final image by predicting the optical flow warp features [13]. In addition, methods [14,15] that generate target results by predicting the object's mask. Although experiments have shown that recent unsupervised methods can perform motion transfer for various objects, achieving accurate motion transfer for relatively non-rigid objects such as the human body is difficult. Due to the human body's complex motion patterns and larger motions, in contrast to more rigid objects like the human face. Therefore, the direct application of motion transfer methods [16] for human bodies often results in problems such as missing semantics and distorted limbs (see Fig. 1 for examples).

Although some subjects can be convincingly animated with existing methods, many things still need improvement. First, existing methods struggle with unclear subject structures and missing semantic parts, which limits their ability to generate real results. Second, the method of warping features through predicting optical flow performs well for small-scale motion, but it is hard to model large-scale and complex motion [17]. Due to the complexity of the human body structure, we believe such methods are not the best choice for solving the human motion transfer task, although they achieve the best performance. Third, the inherent complexity of human motion makes it difficult for the network to compute accurate dense optical flow [18], leading to inaccurate dense optical flow generating unaligned features and ultimately generating unrealistic results.

Accurate semantic information is important in image generation to generate high-quality image results. The key to successful

motion transfer is to find the correspondence of semantic regions between source and driving images [19]. Most previous work has used keypoints features to represent correspondences between bodies, but keypoints do not represent meaningful semantic parts, which represent position and orientation, not shape[5]. Therefore, this may lead to inaccurate motion estimation.

We also noticed the problem of identity confusion and shape interference during the conversion process. Existing methods such as TPS-MM [13] distinguish shape and pose by training an additional network, which may not be effective for complex tasks such as human body motion. Because it can help to recover the shape of the target but is less accurate in recovering the semantic regions of the target human body, which leads to visually implausible situations. For instance, the source figure with a thin body size in the source image may be transformed to the action of a larger body size in the driving image, causing unexpected changes in the source figure's body shape. This is one of the reasons why human motion transfer generalization is very poor.

We aim to fuse the driving image's motion and the source image's identity to generate the target image. To address the above issues, we propose a semantic-guided motion transfer network. Instead of using the correspondence of regions near keypoints, we use pixel-level correspondence, which helps in more accurate motion estimation. First, we extract the semantic features of the driving image and the source image that represent the motion and the identity of the target result, respectively. Second, we adopt a cross-modal attention layer [20] to compute the deformed target feature. It uses the semantic features of the source and driving images as input to learning the correspondence between their semantic regions. Furthermore, to further guide the network to generate the target image, we introduce the rough mask region features of the driving image and use them for the computed target feature and the generation phase. This operation facilitates the decoupling of motion and appearance. Finally, we pass the predicted target features to the generator in the generation phase to produce the target image. In summary, our main contributions are as follows:

1. We innovatively introduce semantic and mask features to model motion and identity. The cross-modal attentional layer guides the network in selecting appropriate feature information to generate more accurate results.
2. We propose a new framework for unsupervised motion transfer. It can be well employed for large and complex human motions and generates accurate semantic structures. Experiments demonstrate that our method outperforms state-of-the-art human image animation methods in motion-related metrics and improves the missing semantics and distorted prevalent during human motion transfer.

2 RELATED WORK

2.1 Motion Transfer & Image Animation

There are currently two main categories of unsupervised motion transfer methods. The first class includes methods that predict optical flow and use it to warp the source image [5,7,8,11,13]. The second category comprises methods that generate target images by predicting the target mask [14,15] or the target warp features [20].

X2Face [11] learns the identity representation of the source image by embedding the network, generating an optical flow to warp the embedded image. Siarohin et al. proposed Monkey-Net [7] and FOMM [8] motion models. Monkey-Net represents motion using sparse keypoints and creates a self-supervised framework for animating arbitrary objects. FOMM [8] further performs a first-order Taylor expansion near each predicted key point, which allows the local affine transform to approximate the motion near each key point more accurately. Although FOMM achieves better results on objects other than faces, it still cannot represent complex motion well because the local affine transform is linear.

In particular, the internal motion of jointed objects such as the human body (e.g., when the arm moves over the body) is not well modeled. The representation of occluded and overlapping parts is still an open problem. To better solve the animation problem of jointed objects, MRAA [5] improves the drawbacks of FOMM by using PCA-based region modeling for better representation of motion and has better quality in representing jointed motion (e.g., human body). However, it is difficult to handle the case where there is a large positional and shape gap between the source and the driving. To better represent complex motions, Zhao et al. [13] analyzed the limitations of MRAA, abandoned the affine transformation, proposed a nonlinear transformation based on the Thin-plate spline to estimate motion, and achieved the best performance. However, similar methods based on explicit motion representation can lead to dramatic degradation of the quality of the generated images due to insufficient prediction of keypoints or regions. These methods tend to perform relatively poorly in cases where there are significant differences in poses and shapes. In contrast, our approach abandons optical flow prediction and instead employs an attention mechanism to learn semantic correspondences between subjects, which can accurately guide the generation process. However, the generalizability of motion transfer remains a major challenge [13].

Toledano et al. [14] estimate motion using a structure mask derived from a keypoint detector based on a motion model. Meanwhile, Shalev et al. [15] proposed a mask-based generator that captures the general pose and shape of the object and uses perturbations to control the source identity of the output frame, reducing the interference of the driving image with the identity information of the source image. Mallya et al. [20] innovatively use a cross-modal attention mechanism to learn the feature correspondence between source and driving images and can be extended to use information from multiple source images to obtain warped feature images implicitly by selecting multiple source features according to the driving images. Although these methods improve some appearance and shape problems during motion transfer, the problem of missing semantics and distortion remains. Our approach borrows from [20] using a cross-modal attention layer, with the difference that we use richer semantic features to represent motion information.

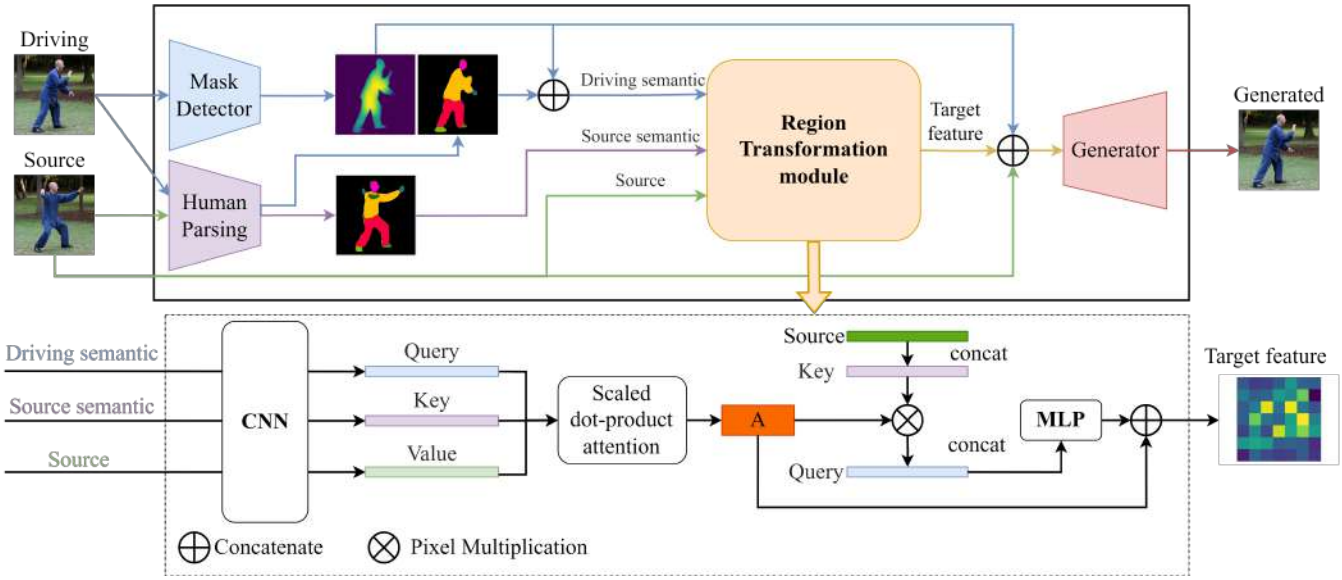

Figure 2. Overview of our approach. The Mask Detector extracts mask features from the driving image. The Human Parsing network extracts semantic features of the foreground character. The Feature Transformation module converts the features of the source frames to the deformed target features. Finally, the target feature is fed into the Generator to generate the final image.

2.2 Attention in Image Animation

Attention-based models [21] were originally used for machine translation tasks. The attention mechanism maps the query and a set of key values to the output. In self-attention, the query, key, and value are derived from the same input feature, and the output is a weighted sum of values. Subsequently, it has also been widely explored and studied in the field of computer vision and used for several computer vision tasks, such as image generation [22], segmentation [23], and image denoising [24]. Moreover, recent works [19,20] have also proposed cross-modal attention, where query, key, and value are obtained from data in two domains.

However, attentional mechanisms remain largely unexplored in human motion transfer. An exception is [20], which uses a cross-modal attention layer where Query and Key are from sparse keypoints for motion representation, while Value is extracted from dense image features. The similarity matrix is obtained through the attention layer, and the features of the source image are warped according to the similarity matrix. Nevertheless, it extracted the key point features that are very sparse, which may lose part of the detail and texture information of the human body. We adopt the same strategy, and the difference is that compared to keypoints, we extract and use richer semantic features and high-resolution features to represent appearance and motion, learning more accurate correspondence of semantic regions for more realistic generated results.

3 METHOD

The overall network architecture of our proposed method is shown in Fig. 2, where the network receives the driving image d and the source image s to generate the final target image ***c***. The method can be described in four parts as follows:

Mask extractor *M* uses a pre-trained MODNet network [25] to extract the subject person's regions. It accepts driving image ***d*** and produces the person's region as the target's regional representation.

Semantic extractor *I* uses the PSPNet network [26] pre-trained for the Single-Human-Parsing task. It accepts the driving image ***d*** and the source image ***s*** and extracts semantic features to represent identity and motion.

Feature transformation module *T* uses a cross-modal attention layer to align the source image features with the driving image. At the same time, the source and driving images are used to recover the features of the target image.

Generator *G* is a two-stage generator network, that concatenates the target features with the driving mask regions and source images, feeding them into the generator network to generate the final target image ***c***.

The network inputs the source frame ***s*** and the driving frame ***d*** to generate the result frame ***c***, where ***c*** contains the subject in the source frame ***s*** and modifies the pose to match the action in the driving frame ***d***. The transformation operation is completed separately for each driving frame in the test phase, and the whole transformation process is as follows:

$$m_d = \boldsymbol{M}(d) \quad (1)$$

$$f_d = \boldsymbol{I}(d) \quad (2)$$

$$f_s = \boldsymbol{I}(s) \quad (3)$$

$$f_t = \boldsymbol{T}(m_d, f_d, f_s, s) \quad (4)$$

$$c = \boldsymbol{G}(m_d\ f_t, s) \quad (5)$$

Here, the feature extractors ***M*** and ***I*** are pre-trained semantic segmentation networks not involved in the network's training. ***M*** uses the MODNet [25] network, which extracts the subject region

of the person driving the image. ***I*** uses the Human Parsing model PSPNet [26] network, which is responsible for extracting the semantic features of the person in the source image. The networks to be trained are the network of obtained *Query*, *Key*, and *Value* in the feature transformation module ***T*** and the generator network ***G***.

3.1 Images Feature Extraction

Previous methods passed the driving and source images through the same network to extract keypoints or mask regions. We need the driving image's motion and the source image's identity, so we hope that the network can extract the motion of the driving image more accurately and the appearance identity information of the source image. Therefore, we extract the semantic features of the source image and the driving image as a model for the appearance and movement of the main character. We also extract the mask region features of the driving character to guide the generation of the target image. Mask region features are useful in image animation tasks. The mask region of the character in the driving image can represent the motion of the driving image since this region is the motion region of the character in the target frame. However, we cannot use a precise target region because this may introduce the problem of identity interference, so we use a rough mask region of the driving image to represent motion. The semantic features of the source image can represent the appearance identity information of the source image better.

First, we utilize region extractor M to extract the region feature m_d of the driving image. The Mask region of the driving subject figure is shown in Fig. 3(a). The shape of the driving image can be recognized and represented as the mask region. However, the same mask region will inevitably introduce the identity information of the driving image and interfere with the generation of the target image. For example, when animating source image A based on video B, the pose of B should be given while discarding the body shape information of B. Otherwise, the generated frames may have the appearance of source A but mixed with the body and edge shape of the person in drive frame B. So we choose the rough target_mask to use as a representation of the motion region of the target image. Where low_hr_pred is the rough result of low pixels obtained from

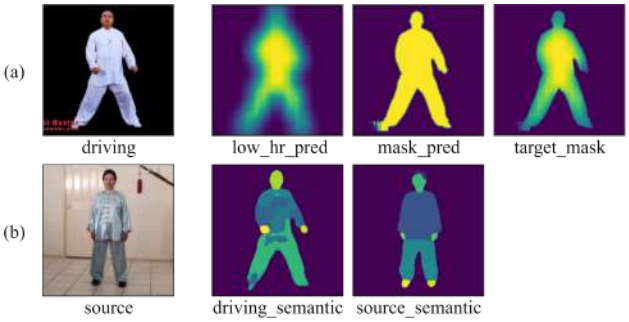

Figure 3. (a) Mask feature visualization. The first column is the driving image. The followings are the low-resolution features, accurate mask features and the target mask feature; (b) Semantic feature visualization, the first column is the source image, followed by the semantic feature visualization results

the network prediction, and the blurring of edges makes it eliminate part of the identity information. Then, multiply it with the accurate prediction *mask_pred* to get *target_mask*. Finally, use a linear interpolation function to scale it to the desired dimension.

Meanwhile, we use the pre-trained Human Parsing model to extract the semantic features f_s of the source image and the semantic features f_s of the driving image. These features capture rich identity information, such as the appearance of the source image. Theoretically, any generic Human Parsing model can extract the features we want. The network outputs a 20-channel vector of the same aspect size as the input image for representing 20 classes of human regions. Here they are directly used as the 20 sub-regions of the human body region, as shown in Fig. 3(b). We take the pre-trained PSPNet network output as the semantic features of the source image, which can well represent the appearance information of the parts of the source image, i.e., the identity representation.

3.1 Feature Transformation

We feed the obtained region feature m_d, the driving frame semantic feature f_d, the source frame semantic feature f_s, and the source image s into the feature transformation module to generate the target feature f_t. m_d and f_d are spliced by an hourglass network to obtain *Query*, and by a similar operation, the *Key* is obtained from f_s, and a downsampling network obtains $\boldsymbol{s}$. We borrowed the hourglass network from [20], but the authors found that the Attention map computed by $Q \times K^T$ is very sparse, so we further reduced the network's depth.

After getting the *Query, Key*, and *Value* after Scaled dot-product attention to get Attention map A, the following equation:

$$Attention(Q,K,V) = softmax\left(\frac{Q\times K^T}{c}\right) \times V = S \times V \quad (6)$$

Here, the vector **S** represents the similarity of **Q**(Query) and **K**(Key), which is multiplied by **V**(Value) after a weighted average. We speculate that this is done to weight the average value of **V** according to the query conditions of **Q** and **K**. This allows the network to choose the appropriate features to reconstruct the image. This is followed by stitching the source image pixels with the learned *Key*, multiplying them with the computed weight matrix **S**. This operation aligns them with the driving image and further stitches them with the query, changing the dimensionality after an MLP layer, and finally obtaining the target features for reconstructing the image, as shown in the bottom half of Fig. 2. The intuition behind using cross-modal attention is that there is a one-to-one spatial correspondence between keys and values, which can learn the semantic region correspondence between the source image and the subject in the driving image. Furthermore, we concatenate image pixels with Key and Query to retain more source image features that can be used to recover information already present in any image. Examples are tilt or rotation, information that may not be easily obtained by averaging weighted values.

3.3 Image Generation

Our approach employs a coarse-to-fine generation strategy during the training and testing phases. Initially, low-resolution

generator ***l*** generates a coarse target image. Subsequently, a high-resolution generator ***h*** is used to generate a more refined target image, where ***h*** can improve the generated image's color and enhance the image's quality.

Specifically, after obtaining the deformed feature f_t, it is fed into the generator **G** with the driving image region mask m_d and the source image ***s*** after stitching and generating the deformed target result. The introduced region feature m_d is mainly used to guide the generator to the region that needs to be repaired. The final image is generated by ***h***.

The input of ***h*** is the output of the first stage concatenated with the source image and the mask of the driving image. So ***h*** can generate the background of the image and part of the detail information of the human body more finely, especially when the pose differences between the source image and the driver image is small. In addition, ***h*** accepts the features of the source image and can recover the identity information of the source image appropriately.

3.3 Training loss

Perceptual loss L_{rec} To train our network, we used the same perceptual loss as in the previous work [13] to train our network. The pre-trained VGG-19 [27] network calculates the reconstruction loss for driving frame ***d*** and generating result ***c*** at different resolutions. The reconstruction loss is written as:

$$L_{rec}(c,\ d) = \sum_j \sum_i \left| V_i(c_j) - V_i(d_j) \right| \quad (7)$$

$V_i(\cdot)$ denotes the output of the i^{th} layer of the VGG-19 network, and j denotes the number of times the image is downsampled to different resolutions. The resolutions are 256×256, 128×128, 64 × 64 and 32 × 32, respectively. The final loss $L_{loss} = \lambda L_{rec}$, where $\lambda = 10$.

4 EXPERIMENT

4.1 Set Up

We train and evaluate our network using the TaichiHD human dataset.

TaiChiHD Dataset is a full-body human action dataset containing hundreds of videos of different motions performed in Tai Chi. As before [8], the videos are resized to 256×256 resolution size and keep the original human scale. There are 2884 training videos and 285 test videos. The dataset has a large range of motion and can better evaluate the model's performance. TaichiHD is the most challenging dataset because it consists of various movements of highly non-rigid bodies [15].

Multiple metrics are used to test our approach. Video reconstruction was used to assess the quality of motion transfer. We used the same metrics as in previous work [5,8,13] for quantitative evaluation and also quantitatively evaluated the generated video's quality. In addition, we also evaluated qualitatively the image animation task, where the source and driver images can be of different identities.

- **Fréchet Inception distance (FID)**: FID evaluates the overall quality of the generated frames, compares the statistical information about the features of the generated frames and the real image, and then calculates the distance between them [28].
- **L1** is the L1 distance between the generated and live ground videos.
- **Average Key-points Distance (AKD)**: AKD measures the average distance between the keypoints of the generated video and the real video using the pre-trained human pose estimator in [29].
- **Missing Key-points Rate (MKR):** MKR detects the percentage of keypoints successfully detected in the real video but missing in the generated video. The output of each keypoint of the human pose estimator [29] indicates whether it was successfully detected.
- **Average Euclidean Distance (AED):** AED measures the average Euclidean distance in feature embedding between the ground truth representation and the generated video [30]. The chosen feature embedding is such that the metric evaluates how well the identity is preserved.

We compared video reconstruction and image animation tasks with three state-of-the-art unsupervised motion transfer methods FOMM [8], TPS-MM [13], and MRAA [5].

4.2 Comparison with Previous Works

4. 2. 1 Video Reconstruction. The benchmark follows the training process because the source and target frames are from the same video. The first frame of the test video is used as the source frame, while the rest of the frames from the same video are used as the driving frames. The goal is to reconstruct all frames of the test video.

Our method shows the quantitative results of FID scores in the last column of Tab.1 and is comparable to those achieved by current methods. Compared to FOMM and MRAA, our results are slightly better but fall short of the latest method TPS-MM. This indicates that our method is close to the state-of-the-art motion transfer methods in terms of overall quality. However, due to the complex background of the images in the TaichiHD dataset, and we did not include modeling of the background, this may be the reason for our slightly poorer results.

The first three columns of Tab. 1 summarize the quantitative results of our video reconstruction. Our method is comparable to state-of-the-art methods. With the best performance in motion-related metrics such as AKD (5% improvement) and MKR (11% improvement), which indicates that our method is more effective in transferring the motion that drives the video. However, our results are slightly worse for other metrics like L1 and AED. L1 may reflect the poorer background of the generated results, while the AED metric is related to the identity of the source image. As shown in Fig. 1, our method can generate more accurate and complete human reconstruction results at larger magnitudes of human motion transfer than the latest method TPS-MM. However, the details in the background and face parts are worse.

More specifically, since we use region and semantic features to model the motion for complex objects like the human body, thus highlighting the integrity of the foreground figure more in the generated results, effectively alleviating the problem of semantic

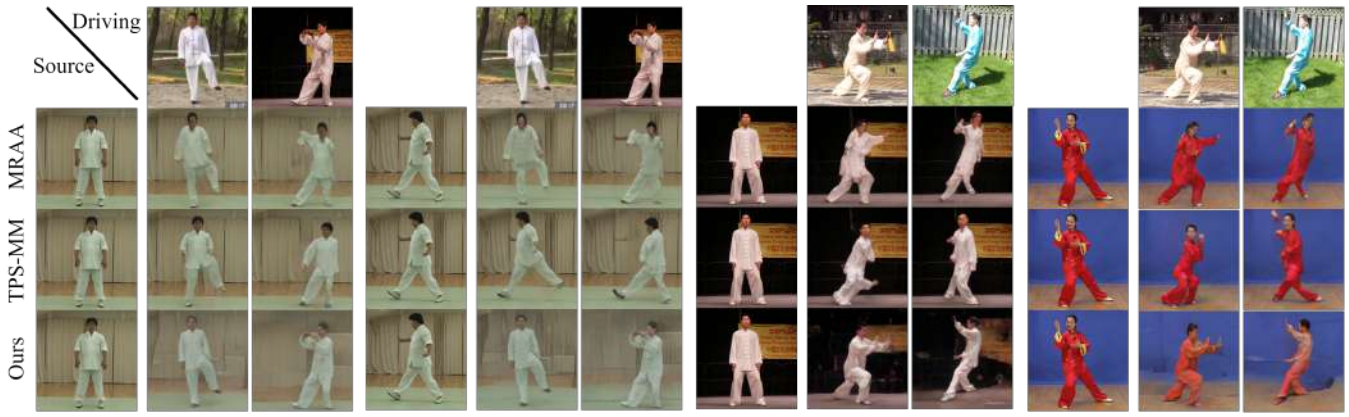

Figure 4. Qualitative comparisons with MRAA [5] and TPS-MM [13] on image animation task. We present the results of four source frame driven by four different frames from the TaichiHD dataset.

Table 1. Quantity results of Video reconstruction. Our results are compared in four benchmark metrics. Lower is better for all metrics, and the best results are in bold.

	TaiChiHD			
Method	L1	(AKD, MKR)	AED	FID
FOMM [8]	0.061	(6.862, 0.036)	0.179	28.08
MRAA [5]	0.048	(5.41, 0.025)	**0.149**	25.74
TPS-MM [9]	**0.045**	(4.57, 0.018)	0.151	**24.29**
OURS	0.049	**(4.34, 0.016)**	0.154	25.57

missing in the generated results, which is reflected in the best AKD and MKD performance. However, we focus more on foreground character motion modeling and appearance rendering but do not consider the background more, which leads to a poorer background in our final results. Additionally, the poorer AED metrics could be due to our rough mask used to eliminate identity interference, which may not have been the optimal choice.

4. 2. 3 Image Animation. We use driving video to animate the source image. The objects in the input source and driving frames have different identities and appearances. Fig. 4 examples of the generated videos for our method and the two current methods. Upon observation, we find that MRAA usually generates bodies that do not match the body shape of the source image or are distorted. While TPS-MM can generate more accurate body movements, it may not capture more detailed semantic information about the body structure, resulting in cases where the semantics of the body parts are missing or distorted (e.g., missing human arms). On the other hand, our approach generally captures motion details of the foreground subject well, resulting in more complete and accurate motion transfer results.

4.3 Ablation

We conducted ablation experiments on the TaichiHD dataset to validate our proposed method. Specifically, we selectively removed different components to analyze their impact on the results. First, we disabled the extracted mask region featured from the driving image and used only the semantic features to estimate motion. Second, we used the full component but removed the high-resolution generator part and used the output of the first stage as the final result. Finally, we used the full framework in training. The quantitative results are shown in Tab. 2, where the first row shows the results without using the features of the mask region of the driving image, and the second row shows the results without using the high-resolution generator. It can be observed that adding region's features of driving images and high-resolution generator improved all metrics.

However, it is worth noting that since the method is purely data-driven, it may not perform well for extreme cases that do not exist in the training dataset. For example, the generation of the character's face may fail during the transfer of the character from the back to the front. Moreover, our method has some limitations in preserving the identity of the source image and facing complex backgrounds.

Table 2. Ablation analysis on the reconstruction task for TaiChiHD. (Lower is better, best result in bold)

	L1	**(AKD, MKR)**	**AED**
No mask feature	0.059	(4.56, 0.017)	0.161
No high generator	0.057	(4.34, 0.016)	0.157
Full Model	**0.049**	**(4.34, 0.016)**	**0.154**

5 CONCLUSION

In this paper, we first discuss the shortcomings of previous work on motion transfer using predicted optical flow warp features to generate results. The main drawback is that it tends to missing semantics and unclear structure when dealing with complex motions such as the human body. To this end, We propose a novel human motion transfer approach that leverages rich semantic features and mask regions to effectively transfer motion and reconstruct appearance. Instead of predicting optical flow, we employ a cross-modal attention approach to compute target features

and use coarse mask region features to guide the initial results generation for target features. Second, we adopt a coarse-to-fine generation strategy using two generators to improve the detail fidelity and quality of the results. Finally, we demonstrate experimentally that our approach performs well in benchmark tests and improves motion-related metrics. It effectively improves problems such as missing semantics and distorted common in the human body during large motion transfer.

REFERENCES

[1] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. 2021. One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing. Retrieved February 10, 2023 from http://arxiv.org/abs/2011.15126

[2] Zhaoying Pan and Jinge Ma. 2022. Face Animation with Multiple Source Images. Retrieved December 11, 2022 from http://arxiv.org/abs/2212.00256

[3] Maxime Oquab, Pierre Stock, Oran Gafni, Daniel Haziza, Tao Xu, Peizhao Zhang, Onur Celebi, Yana Hasson, Patrick Labatut, Bobo Bose-Kolanu, Thibault Peyronel, and Camille Couprie. 2021. Low Bandwidth Video-Chat Compression using Deep Generative Models. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Nashville, TN, USA, 2388–2397. DOI:https://doi.org/10.1109/CVPRW53098.2021.00271

[4] Borun Xu, Biao Wang, Jiale Tao, Tiezheng Ge, Yuning Jiang, Wen Li, and Lixin Duan. 2021. Move As You Like: Image Animation in E-Commerce Scenario. In *Proceedings of the 29th ACM International Conference on Multimedia*, ACM, Virtual Event China, 2759–2761. DOI:https://doi.org/10.1145/3474085.3478550

[5] Aliaksandr Siarohin, Oliver J. Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. 2021. Motion Representations for Articulated Animation. Retrieved November 12, 2022 from http://arxiv.org/abs/2104.11280

[6] Subin Jeon, Seonghyeon Nam, Seoung Wug Oh, and Seon Joo Kim. 2020. Cross-Identity Motion Transfer for Arbitrary Objects through Pose-Attentive Video Reassembling. Retrieved November 17, 2022 from http://arxiv.org/abs/2007.08786

[7] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. Animating Arbitrary Objects via Deep Motion Transfer. Retrieved November 12, 2022 from http://arxiv.org/abs/1812.08861

[8] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2020. First Order Motion Model for Image Animation. Retrieved November 12, 2022 from http://arxiv.org/abs/2003.00196

[9] Jiale Tao, Biao Wang, Tiezheng Ge, Yuning Jiang, Wen Li, and Lixin Duan. 2022. Motion Transformer for Unsupervised Image Animation. Retrieved November 12, 2022 from http://arxiv.org/abs/2209.14024

[10] Jilin Tang, Yi Yuan, Tianjia Shao, Yong Liu, Mengmeng Wang, and Kun Zhou. 2021. Structure-aware Person Image Generation with Pose Decomposition and Semantic Correlation. Retrieved November 13, 2022 from http://arxiv.org/abs/2102.02972

[11] Olivia Wiles, A. Sophia Koepke, and Andrew Zisserman. 2018. X2Face: A Network for Controlling Face Generation Using Images, Audio, and Pose Codes. In *Computer Vision – ECCV 2018*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu and Yair Weiss (eds.). Springer International Publishing, Cham, 690–706. DOI:https://doi.org/10.1007/978-3-030-01261-8_41

[12] Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. 2019. DwNet: Dense warp-based network for pose-guided human video generation. Retrieved November 12, 2022 from http://arxiv.org/abs/1910.09139

[13] Jian Zhao and Hui Zhang. 2022. Thin-Plate Spline Motion Model for Image Animation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, New Orleans, LA, USA, 3647–3656. DOI:https://doi.org/10.1109/CVPR52688.2022.00364

[14] Or Toledano, Yanir Marmor, and Dov Gertz. 2021. Image Animation with Keypoint Mask. DOI:https://doi.org/10.13140/RG.2.2.16342.16968

[15] Yoav Shalev and Lior Wolf. 2022. Image Animation with Perturbed Masks. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, New Orleans, LA, USA, 3637–3646. DOI:https://doi.org/10.1109/CVPR52688.2022.00363

[16] Borun Xu, Biao Wang, Jinhong Deng, Jiale Tao, Tiezheng Ge, Yuning Jiang, Wen Li, and Lixin Duan. 2022. Motion and Appearance Adaptation for Cross-domain Motion Transfer. In *Computer Vision – ECCV 2022*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella and Tal Hassner (eds.). Springer Nature Switzerland, Cham, 529–545. DOI:https://doi.org/10.1007/978-3-031-19787-1_30

[17] Jiale Tao, Biao Wang, Borun Xu, Tiezheng Ge, Yuning Jiang, Wen Li, and Lixin Duan. 2022. Structure-Aware Motion Transfer with Deformable Anchor Model. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, New Orleans, LA, USA, 3627–3636. DOI:https://doi.org/10.1109/CVPR52688.2022.00362

[18] Zhimeng Zhang and Yu Ding. 2022. Adaptive Affine Transformation: A Simple and Effective Operation for Spatial Misaligned Image Generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, ACM, Lisboa Portugal, 1167–1176. DOI:https://doi.org/10.1145/3503161.3548330

[19] Evan Casey, Victor Perez, and Zhuoru Li. 2021. The Animation Transformer: Visual Correspondence via Segment Matching. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Montreal, QC, Canada, 11303–11312. DOI:https://doi.org/10.1109/ICCV48922.2021.01113

[20] Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. 2022. Implicit Warping for Animation with Image Sets. Retrieved November 12, 2022 from http://arxiv.org/abs/2210.01794

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. Retrieved April 3, 2023 from http://arxiv.org/abs/1706.03762

[22] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. 2023. All are Worth Words: A ViT Backbone for Diffusion Models. Retrieved April 18, 2023 from http://arxiv.org/abs/2209.12152

[23] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. 2022. GroupViT: Semantic Segmentation Emerges from Text Supervision. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, New Orleans, LA, USA, 18113–18123. DOI:https://doi.org/10.1109/CVPR52688.2022.01760

[24] Juncheng Li, Bodong Cheng, Ying Chen, Guangwei Gao, and Tieyong Zeng. 2023. EWT: Efficient Wavelet-Transformer for Single Image Denoising. Retrieved April 18, 2023 from http://arxiv.org/abs/2304.06274

[25] Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson W. H. Lau. 2022. MODNet: Real-Time Trimap-Free Portrait Matting via Objective Decomposition. Retrieved April 6, 2023 from http://arxiv.org/abs/2011.11961

[26] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. 2019. Self-Correction for Human Parsing. Retrieved November 20, 2022 from http://arxiv.org/abs/1910.09777

[27] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. Retrieved April 19, 2023 from http://arxiv.org/abs/1603.08155

[28] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2018. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. Retrieved April 6, 2023 from http://arxiv.org/abs/1706.08500

[29] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. Retrieved April 6, 2023 from http://arxiv.org/abs/1611.08050

[30] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In Defense of the Triplet Loss for Person Re-Identification. Retrieved April 6, 2023 from http://arxiv.org/abs/1703.07737