Motion and Appearance Adaptation for Cross-Domain Motion Transfer

Borun Xu[1], Biao Wang[2], Jinhong Deng[1], Jiale Tao[1], Tiezheng Ge[2], Yuning Jiang[2],
Wen Li[1*], and Lixin Duan[1]

[1] University of Electronic Science and Technology of China
[2] Alibaba Group
xbr_2017@std.uestc.edu.cn,
{jhdeng1997, jialetao.std, liwenbnu, lxduan}@gmail.com,
{eric.wb, tiezheng.gtz, mengzhu.jyn}@alibaba-inc.com

Abstract. Motion transfer aims to transfer the motion of a driving video to a source image. When there are considerable differences between object in the driving video and that in the source image, traditional single domain motion transfer approaches often produce notable artifacts; for example, the synthesized image may fail to preserve the human shape of the source image (*cf*. Fig. 1 (a)). To address this issue, in this work, we propose a Motion and Appearance Adaptation (MAA) approach for cross-domain motion transfer, in which we regularize the object in the synthesized image to capture the motion of the object in the driving frame, while still preserving the shape and appearance of the object in the source image. On one hand, considering the object shapes of the synthesized image and the driving frame might be different, we design a shape-invariant motion adaptation module that enforces the consistency of the angles of object parts in two images to capture the motion information. On the other hand, we introduce a structure-guided appearance consistency module designed to regularize the similarity between the corresponding patches of the synthesized image and the source image without affecting the learned motion in the synthesized image. Our proposed MAA model can be trained in an end-to-end manner with a cyclic reconstruction loss, and ultimately produces a satisfactory motion transfer result (*cf*. Fig. 1 (b)). We conduct extensive experiments on human dancing dataset Mixamo-Video to Fashion-Video and human face dataset Vox-Celeb to Cufs; on both of these, our MAA model outperforms existing methods both quantitatively and qualitatively.

1 Introduction

Given a source image and a driving video of the same object, motion transfer (*a.k.a.* image animation) aims to generate a synthesized video that mimics the motion of the driving video while preserving the appearance of the source image. It recently received increasing attention, due to its potential applications in real-world scenarios, such as face swapping[38,31,39], dance transferring[5], *etc*.

[*] The corresponding author

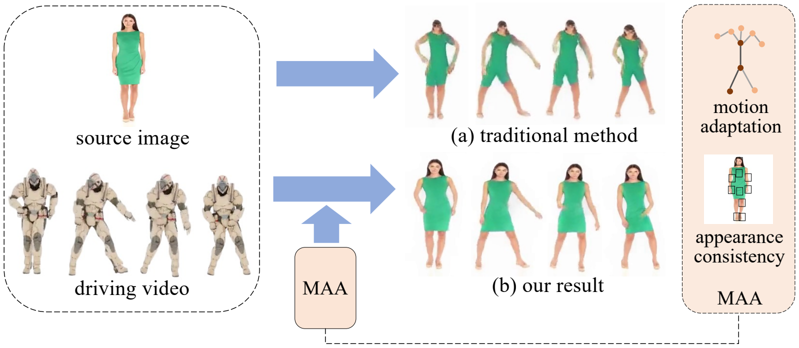

Fig. 1: Motion transfer results. (a) is generated by traditional motion transfer model trained on source domain videos only and (b) is generated by our proposed MAA model

Many works in this field focus on the single-domain motion transfer [30,31,32], where the driving video and source image come from the same domain. However, in real applications, there are often requirements to transfer motion among different domains. For example, as shown in Fig 1, the e-commerce companies might be interested in animating a fashion model to attract consumers by learning the robot dance from a Mixamo character. However, due to the differences in shape and cloth between the Mixamo character and the fashion model, traditional single-domain motion transfer approaches often produce notable artifacts in the synthesized image (*e.g.*, failing to preserve the human shape of the fashion model (*cf*. Fig. 1 (a))).

To this end, in the present work, we study the cross-domain motion transfer problem and propose a novel Motion and Appearance Adaptation (MAA) approach to address this issue. Specifically, traditional motion transfer methods usually take two arbitrary frames of the same video as source image and driving frame for learning motion with a reconstruction loss, because the two frames share the same appearance and shape. However, such training mode cannot be directly applied to the cross-domain motion transfer because no ground-truth is available. In our proposed MAA approach, we build a cyclic reconstruction pipeline inspired by CycleGAN[42] and cross-identity[19]. In particular, given a source image and a driving frame obtained from different domains, we first obtain a synthesized image using a basic motion transfer (MT) model, *e.g.*, the model in[31] or[32]. We next arbitrarily take another frame from the driving video as the source image and the synthesized image as a driving frame, and input them into the basic MT model to produce the second synthesized image. Because the second synthesized image should mimic both the motion and appearance of original driving frame, a cyclic reconstruction loss can be applied for training. In this way, we obtain a motion transfer model for cross-domain motion transfer.

Moreover, since the source image and driving frame are drawn from different domains, while the topology of the object structure (*e.g.*, the skeleton) is similar, the configurations of the object structure (*e.g.*, the human body shape) often deviate. When doing motion transfer, we should be aware of such difference and keep the object shape of synthesized image be similar to the source image while unaffected by the driving frame. For this purpose, we design a shape-invariant motion adaptation module and a structure-guided appearance consistency module to regularize the basic motion transfer model.

Specifically, in the shape-invariant motion adaptation module, we design an angle consistency loss to enforce the angles of the corresponding object parts in the synthesized image to be similar to those of the driving frame, such that the motion of this frame can be mimicked well without changing the object shape. In the structure-guided appearance consistency module, we extract image patches from the synthesized image and the source images based on the object structure and enforce the corresponding patches to be similar; this ensures that the appearance of the synthesized image and the source image are consistent, even though the motions of the two images are different.

The entire process can be trained in an end-to-end manner, and finally our MAA model can effectively perform motion transfer across domains while also properly preserving the shape and appearance of the object (*cf*. Fig. 1 (b)). We validate our proposed approach on two pairs of datasets: the human body datasets Mixamo-Video to Fashion-Video[40] and the human face datasets Vox-Celeb[28] to Cufs[37]. Extensive experimental results demonstrate the effectiveness of our proposed approach. Our source code will be released soon.

2 Related Work

Motion Transfer: Current motion transfer approaches can be categorized into two types: model-based and model-free approaches. The model-based approaches mainly focus on human body pose transfer[26,27,3], which utilize a pre-trained pose estimator or key point detector to extract the pose of driving image as a guidance information. And a number of researchers followed such setting[23,29,43,22,16,20]. Moreover, a series of works apply this model-based pattern on human facial expression transfer[4,7,13]. Like body pose transfer, they also employ a pre-trained facial landmark detector to model the facial expression.

The model-free approaches[30,31,19,32,34] does not rely on pre-trained third-party models, and extend the model-based method to arbitrary objects. Aliaksandr *et al.* [30] proposed a model-free motion transfer model Monky-Net that can apply motion transfer on arbitrary objects with an unsupervised key point detector trained by reconstruction loss [18]. Aliaksandr *et al.* [31] further improved Monkey-Net to FOMM to solve the large motion problem. The unsupervised key point detector is also utilized in FOMM, with local affine transformations being added for motion modeling. A generator module is utilized to generate final result with the warped source image feature. Subin *et al.* [19] proposed pose attention mechanism with an unsupervised key point detector to model motion. Recently, Aliaksandr *et al.* [32] improved FOMM with an advanced motion model and background motion model to MRAA. Although promising results are achieved for the single domain motion transfer, these methods might suffer from performance degradation when the source image and driving video come from different domains, where a considerable appearance difference often exists. Recently, Wang *et al.* [36] used encoder based motion transfer approach which can be applied to the cross-domain scenario, and better results are achieved compared with the single domain motion transfer Monkey-Net model. In contrast, our proposed MAA approach is a general framework, and can integrate traditional motion transfer model like FOMM and MRAA to produce excellent results for large motion.

Domain Adaptation: Many works have been proposed to handle the scenario where the training and test data comes from different domains for different computer vision tasks , *e.g.*, classification [8], semantic segmentation [25,24,10,11], object detection [6,9], pose estimation[21,41], *etc*. A majority of works were developed to learn domain-invariant features using the domain adversarial learning [35,12]. Cross-domain motion transfer is more complicated, since we need to capture motion from the the driving video while preserving the appearance from the source domain. Nevertheless, the strategies proposed in traditional domain adaptation works might be useful to help motion transfer. For example, we apply the cyclic training pipeline inspired by CycleGAN [42], and build our patch-based appearance consistency module based on PatchGAN [17].

3 Methodology

In this section, we present our Motion and Appearance Adaptation approach for cross-domain motion transfer. Formally, let us denote a driving video as $V_d = \{I_d^i|_{i=1}^T\}$, where each I_d^i is a driving frame, while a source image is denoted as I_s; thus, the task of motion transfer is to synthesize a new video $\hat{V}_d = \{\hat{I}_d^i|_{i=1}^T\}$, where each $\hat{I}_d$ adequately captures the object motion in the corresponding driving frame I_d^i while also preserving the object appearance of the source image I_s.

The appearance of an object roughly consists of two aspects, *shape* and *texture*. The shape largely refers to its geometric property (*e.g.*, length, slimness, etc.), while the texture usually means how the object looks like regardless of its shape (*e.g.*, dresses with different colors). Traditional motion transfer methods generally assume that the driving frame and the source image are derived from the same domain, where they implicitly suppose the object shapes are similar. Consequently, when the source image is derived from a new domain with different object shapes, these methods often fail to preserve the shape of the object in the source image.

In this work, we study the cross-domain motion transfer problem, in which the source image and driving frame are from different domains. In other words, there might be considerable differences in appearance between them in terms of both shape and texture. An example is given in Fig. 1, where both the clothes and body shapes of the fashion model and the Mixamo character exhibit notable differences.

In what follows, we first present an overview of the pipeline of our proposed MAA approach in Sec. 3.1, after which we present the shape-invariant motion adaptation (SIMA) module and structure-guided appearance consistency (SGAC) module in Sec. 3.2 and Sec. 3.3 respectively; these effectively learning the motion and appearance from the driving frame and source image, respectively.

3.1 Overview

We design a cyclic training pipeline for cross-domain motion transfer, as shown in the right-hand part of Fig. 2. The pipeline consists of a basic motion transfer model, our

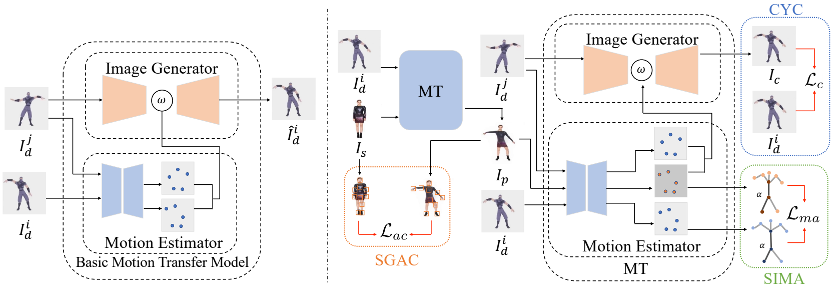

Fig. 2: The pipeline of our proposed method. The left-hand side is the architecture of the traditional single domain motion transfer model FOMM[31], which is used as a basic motion transfer model in our approach. Moreover, the right-hand side is the framework of our proposed MAA method where we design a cyclic reconstruction loss (CYC), a shape-invariant motion adaptation (SIMA) and a structure-guided appearance consistency (SGAC) module

proposed shape-invariant motion adaptation module and structure-guided appearance consistency module, and a cyclic loss.

Basic Motion Transfer Model: The basic motion transfer (MT) model follows the traditional motion transfer model[31,32]. We here illustrate the basic MT model by taking FOMM[31] as an example, and other models like[32] can be similarly integrated into our pipeline.

As shown in the left-hand part of Fig. 2, traditional motion transfer methods typically employ a reconstruction training mode for learning and synthesizing motion. During the training phase, they select two arbitrary frames from the driving video as the source image and driving frame, which are used as input of the MT model. For each image, the motion keypoints and their local affine transformation are extracted using a motion estimator, where the motion keypoints can be conceptualized as the centroids of moving object parts. The dense motion flow from the source image to the driving frame can therefore be estimated using their motion keypoints and affine transformations. In the next step, the dense motion flow is used to warp the feature map of the source image, and produce the synthesized image $\hat{I}_d^i$ using the image generator. A perceptual loss is used as the reconstruction loss after the image generator to ensure that the synthesized image $\hat{I}_d^i$ fully reconstructs the driving frame I_d^i as in[31]:

$$\mathcal{L}_r = \sum_{l=k}^{K} |F_l(\hat{I}_d^i) - F_l(I_d^i)| \qquad (1)$$

where $F_l(\cdot)$ is feature map output by the l-th layer of a pre-trained VGG-19 network[33].

Researchers have proposed different method[32] to improve the motion estimator in order to more precisely extract motion information, yet the motion representation (*i.e.*, motion keypoints and affine transformations) remains similar. In the interests of simplicity, we depict only the motion keypoints in Fig. 2, which are related to our MAA approach. Readers can refer to[31] for further details.

Cyclic Training Pipeline: In cross-domain motion transfer, the source image and driving video are obtained from different domains. So it is undesirable to pick a frame in the driving video as a source image and apply the reconstruction loss after the image generator, as the model will inevitably be overfitted to the driving video, which will lead to artifacts in the synthesized image.

To address this issue, we build a cyclic reconstruction framework inspired by the CycleGAN[42] and cross-identity[19]. As shown in the right-hand side of Fig 2, we employ two basic basic MT models that share the same parameters. Given a source image I_s and a driving frame I_d^i, we first obtain a synthesized image I_p using the basic MT model. Since there is no ground-truth for the synthesized image, the reconstruction loss cannot be used for I_p.

We then take the synthesized image I_p as a driving frame, along with an arbitrary frame I_d^j as the source image, and these are input again into the basic MT model to produce another synthesized image I_c. Intuitively, I_c should mimic the motion of I_p, as well as I_d^i, since we expect I_p to mimic the motion of I_d^i. At the same time, I_c should also preserve the appearance of I_d^j, as well as I_d^i, which is derived from the same driving video as I_d^j. This allows us to employ I_d^i and the cyclically generated I_c to create a reconstruction loss for training. More specifically, we employ the perceptual loss similarly as in Eq. (1):

$$\mathcal{L}_c = \sum_{l=k}^{K} |F_l(I_c) - F_l(I_d^i)| \tag{2}$$

While the cyclic reconstruction loss enables us to train the motion transfer model in the cross-domain setting, this is only a weak supervision that cannot fully guarantee a satisfactory result. We therefore further introduce the shape-invariant motion adaptation module and patch-based appearance model to regularize the motion transfer process, which will be explained in more detail below.

3.2 Shape-invariant Motion Adaptation

Due to the significant appearance difference between the source image I_s and the driving frame I_d, the generated synthesized image I_p often fails to adequately capture the object motion in the driving frame I_d. We therefore propose to directly regularize the object pose in I_p with that in I_d based on the extracted motion keypoints.

However, due to the diversity of the object shapes in I_p and I_d, it is undesirable to directly regularize the consistency of their keypoint positions. We therefore propose to discover the intrinsic topology of the object, then regularize the included angles between adjacent object parts of two objects.

Structure Topology Discovery: To discover the intrinsic object topology, for each driving video, we employ a pre-trained basic MT model to extract the motion keypoints of all frames in the video. Because the motion keypoints roughly describe the objects' moving body parts, two keypoints can be considered to be adjacent if their distance does not change substantially between different frames.

Formally, given a driving frame I_d, we denote its motion keypoints as $\mathcal{K}_d = \{\mathbf{k}_d^i|_{i=1}^K\}$, where K is the number of motion keypoints. For each pair of keypoints $\mathbf{k}_d^i$ and $\mathbf{k}_d^j$, we

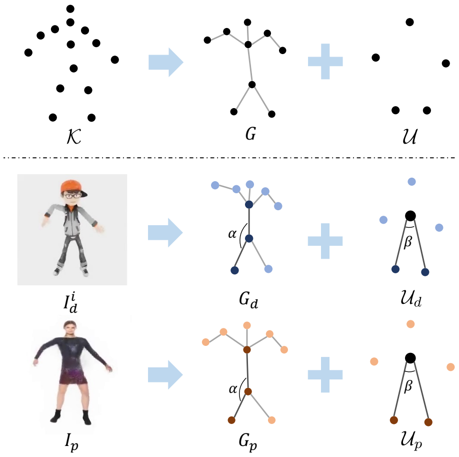

Fig. 3: Illustration of our shape-invariant motion adaptation module. The top row show the structure topology, and the bottom two rows represents the motion adaptation stage using structured and unstructured keypoints

calculate their ℓ_2 distance $d_{i,j} = \ell_2(\mathbf{k}_d^i, \mathbf{k}_d^j)$, where $i \neq j$. The average distance across all frames of all driving videos can then be computed as $\bar{d}_{i,j} = \frac{1}{T}\sum_{t=1}^{T} d_{i,j}^{(t)}$, where $d_{i,j}^{(t)}$ is the distance in the t-th frame, while T is the total number of video frames. Finally, we calculate the total distance diversity of $\mathbf{k}_d^i$ and $\mathbf{k}_d^j$ as follows:

$$v_{i,j} = \sum_{t=1}^{T} |d_{i,j}^{(t)} - \bar{d}_{i,j}| \tag{3}$$

Intuitively, the distance diversity describes the stability of the connection between two keypoints. The smaller the distance diversity $v_{i,j}$, the higher the likelihood that the two keypoints will be adjacent. We then use the distance diversities to construct a structure topology graph G, where the nodes are keypoints, and the edges are defined according to the distance diversities. Specifically, we define the edge value as follows:

$$e_{i,j} = \begin{cases} \frac{(v_{i,j}-\eta)^2}{\eta^2}, & v_{i,j} < \eta, \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where η is a threshold, and we filter out the edges with high distance diversities, as these imply that the two keypoints are unlikely to be adjacent. Note that the edge value $e_{i,j}$ is within the range of $[0, 1]$. It can be seen as a measurement of the strength of the connection between two keypoints. We will demonstrate that it can also be used as a weight when we regularize the keypoints between driving frame and synthesized image.

Moreover, it is possible that not all keypoints are connected in a single graph; we select the largest graph as our structure topology graph G. We refer to the keypoints in G as the structured keypoints and the others as unstructured keypoints. For improved convenience of presentation, we denote the set of structured keypoints as $\mathcal{S}$ and their edges as $\mathcal{E}$, the structure topology graph can be presented as $G = \{\mathcal{S}, \mathcal{E}\}$. For unstructured keypoints, we retain only the keypoints and discard their edges, since their connectivities are weak, and denote the set of unstructured keypoints as $\mathcal{U}$. We present an illustration of the structure topology discovery in the top row of Fig. 3.

Regularizing Structured Keypoints: Given a driving frame I_d and the synthesized I_p, we extract their keypoints $\mathcal{K}_d = \{\mathbf{k}_d\}$ and $\mathcal{K}_p = \{\mathbf{k}_p\}$ using the basic MT model. To regularize the keypoints in the driving frame I_d and the synthesized I_p, we instantiate the structure topology G using the extracted keypoints $\mathcal{K}_d$ and $\mathcal{K}_p$, respectively. Taking the driving frame as an example, the instantiated graph is presented as $G_d = \{\mathcal{S}_d, \mathcal{E}_d\}$; here, $\mathcal{S}_d$ is the set of structured keypoints in I_d, while $\mathcal{E}_d$ is the set of corresponding edges which are calculated based on the Euclidean distances between keypoints. The instantiated graph of the synthesized image $G_p = \{\mathcal{S}_d, \mathcal{E}_d\}$ can be similarly defined. We illustrate the instantiated graphs in the top of Fig 3.

When examining the structured keypoints, we can observe that considerable differences exist in terms of object shape; this validates our analysis that it is not preferable to directly regularize the keypoint positions. However, the pose can be portrayed as the included angle of each triplet of the connected keypoints in the structure graph.

Specifically, taking the driving frame as an example, let us define a triplet of connected keypoints as $\mathbf{t}_d = \{\mathbf{k}_d^i, \mathbf{k}_d^j, \mathbf{k}_d^k\}$, where both $\mathbf{k}_d^j$ and $\mathbf{k}_d^k$ are connected to $\mathbf{k}_d^i$. We further denote the set of all keypoint triplets in G_d as $\mathcal{T}_d = \{\mathbf{t}_d^n|_{n=1}^N\}$, where N is the total number of triplets. Similarly, we define the set of triplets for the synthesized image as $\mathcal{T}_p = \{\mathbf{t}_p^n|_{n=1}^N\}$.

For each triplet $\mathbf{t}_d^n$ (*resp.*, $\mathbf{t}_p^n$), we calculate its included angle and denote it by $\alpha(\mathbf{t}_d^n)$ (*resp.*, $\alpha(\mathbf{t}_p^n)$). We then regularize the consistency of the corresponding included angles for structured keypoints in the driving frame and the synthesized image as follows:

$$\mathcal{L}_{rs} = \frac{1}{N}\sum_{n=1}^{N} \gamma_n |\alpha(\mathbf{t}_d^n) - \alpha(\mathbf{t}_p^n)| \tag{5}$$

where γ_n is the weight for the n-th triplet. We calculate γ_n using the edge values in the topology graph G. Specifically, given any triplet $\mathbf{t} = \{\mathbf{k}^i, \mathbf{k}^j, \mathbf{k}^k\}$ in the topology graph, the weight is computed as $\gamma = e_{i,j}e_{i,k}$. As the edge represents the strength of the connections between two keypoints, it is reasonable to employ the multiplication of the two edges that formed the included angle as the weight for regularization.

Regularizing Unstructured Keypoints: Similarly, given a driving frame I_d and the synthesized I_p, we identify their unstructured keypoints $\mathcal{U}_d$ and $\mathcal{U}_p$, respectively. Since these unstructured keypoints are disjoint, we constrain them by encoding their included angles with the object centroid. Taking the driving frame as an example, for each pair of keypoints $(\mathbf{k}_d^i, \mathbf{k}_d^j)$ in $\mathcal{U}_d$, we construct a triplet $\hat{\mathbf{t}}_d = (\mathbf{k}_d^i, \mathbf{k}_d^c, \mathbf{k}_d^j)$ in which $\mathbf{k}_d^c$ is the object centroid, and further denote the included angle as $\beta(\hat{\mathbf{t}}_d)$. We similarly define the corresponding included angle for the synthesized image as $\beta(\hat{\mathbf{t}}_p)$. We then regularize the consistency of the corresponding included angles for the structured keypoints in the driving frame and the synthesized image as follows:

$$\mathcal{L}_{ru} = \frac{1}{\hat{N}}\sum_{n=1}^{\hat{N}} |\beta(\hat{\mathbf{t}}_d^n) - \beta(\hat{\mathbf{t}}_p^n)| \tag{6}$$

where $\hat{N}$ is the number of constructed triplets using structured keypoints in each image.

Combining the loss of structured and unstructured keypoints, the total loss of our shape-invariant motion adaptation loss can be written as follows:

$$L_{ma} = L_{rs} + L_{ru} \tag{7}$$

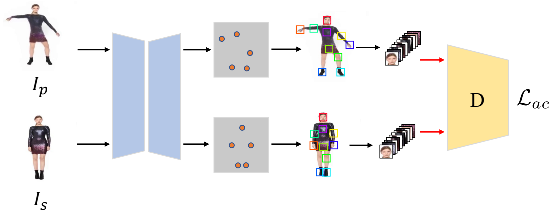

Fig. 4: Illustration of our structure-guided appearance consistency module

3.3 Structure-Guided Appearance Consistency Module

We now consider how the appearance of the synthesized image I_p might be enforced to be similar to that of the source image I_s. Note that the object poses in I_p and I_s are different, as we have enforced I_p to mimic the pose of the driving frame. We therefore propose structure-guided appearance consistency module to regularize the appearance consistency of object parts in I_p and I_c to avoid impacting the learned object pose of I_p

In particular, we use the predicted motion keypoints to extract image patches of fixed size from both images. After collecting the patches from I_p (*resp.*, I_s), a discriminator $\mathcal{D}$ is then introduced to enforce the appearance consistency between the corresponding patches by means of an adversarial training strategy, as shown in Fig. 4. The aim of the discriminator is to determine whether the input patches are from I_p or I_s by minimizing a cross-entropy loss, while the generation model $\mathcal{B}$ (*i.e.*, the basic MT model) aims at generating pseudo-images $\mathcal{B}(I_s)$, which are difficult to distinguish from the source image I_s by maximizing the cross-entropy loss. Formally, we express the loss of our patch-based appearance consistency module as follows:

$$L_{ac} = \log \mathcal{D}(V(I_s)) + \log(1 - \mathcal{D}(V(\mathcal{B}(I_s)))) \tag{8}$$

where $V(\cdot)$ represents the patch extraction operation.

3.4 Summary

We combine all losses together to train our proposed MAA model in an end-to-end manner. The overall objective function can be written as follows,

$$\mathcal{L} = \mathcal{L}_r + \mathcal{L}_c + \lambda_{ma}\mathcal{L}_{ma} - \lambda_{ac}L_{ac} \tag{9}$$

where λ_{ma} and λ_{ac} are tradeoff parameters used to balance the losses. Due to the existence of the discriminator, we optimize the overall loss in an adversarial training manner, *i.e.*, $\min_{\mathcal{B}} \max_{\mathcal{D}} \mathcal{L}$. Detailed training loop is presented in Supplementary materials.

4 Experiment

4.1 Datasets

We conduct experiments for two types of object including human body and human face. For the human body animation, we transfer motion from Mixamo-Video to Fashion-Video, and for human face animation we transfer motion from Vox-Celeb to CUHK Face Sketch (Cufs).

Mixamo-Video Dataset is a synthetic human dancing video dataset newly constructed by ourselves. We collect 15 characters of 3D human body models and 46 dancing sequences from the mixamo [1] website, then render dancing videos for these characters and dancing sequences, leading to $15 \times 46 = 690$ videos in total with resolution of 256×256. We split ten of the characters as training set and the rest as test set, *i.e*. 460 and 230 videos, respectively. Details of the dataset are presented in supplementary materials, and we will release the dataset soon.

Fashion-Video Dataset is a video dataset for showing clothes. It contains 500 training videos and 100 testing videos with size of 256×256.Although it is a video dataset, We use it as an image dataset by selecting one frame per video randomly in training stage.

Vox-Celeb is a video dataset of human talking. It consists of $12,331$ training videos and 444 testing videos resized to 256×256.

CUHK Face Sketch (Cufs) is an image dataset of human face sketches. The dataset contains 305 images where training set and test set have 250 and 45 images *resp.*. Each image is a sketch drawn by an artist based on a photo taken in a frontal pose with a natural expression. We also resize those images into the size of 256×256.

Table 1: Quantity results comparison of our method with source only FOMM model and MRAA model. The lower FID and AED values are the better

	Mixamo ⟶ Fashion		Vox ⟶ Cufs	
	FID ↓	AED ↓	FID ↓	AED ↓
MRAA	177.3	0.376	127.1	0.764
FOMM	175.9	0.359	112.5	0.693
Ours (MRAA)	72.1	0.289	86.5	0.627
Ours (FOMM)	**61.7**	**0.274**	**50.1**	**0.573**

4.2 Quantitative Results

Metrics: As the ground-truth video are not available in cross-domain motion transfer, to quantitatively assess the synthesized videos, we employ two metrics for evaluate generative models as follows

- **Fréchet Inception distance (FID)[15]** This score indicates the overall quality of generated frames, it compares the feature statistics of generated frames and real images, then calculates the distance between them.

- **Average Euclidean Distance (AED)**[31] Considering the generated images share the same identity with source images, we utilize AED to evaluate the identity similarity between them. It also computes the feature distance between two input images. Specifically, a pre-trained person re-identification network [14] and a pre-trained facial identification network [2] are used to extract identity feature representations for human body and human face dataset, respectively.

Results: As aforementioned, unsupervised motion transfer models like FOMM[31] or MRAA[32] can be integrated into our MAA framework as the basic motion transfer model. We conduct experiments by respectively using FOMM and MRAA as our basic motion transfer model, and take the original FOMM and MRAA as the corresponding baseline for comparison. For both methods, the baseline models are trained on the driving video dataset without considering the cross-domain issue. Note that the newly proposed modules in our MAA model are only used in training stage, and the model in the test stage share the same architecture with the baseline FOMM or MRAA model.

We report the results for Mixamo-Video → Fashion-Video and Vox → Cufs in Tab. 1. Comparing with the FOMM and MRAA model, our proposed MAA approach achieves a much better performance. In particular, compared with FOMM, we achieve a FID of 61.7 and an AED of 0.274 for Mixamo-Video → Fashion-Video, and 50.1 *vs*. 112.5 and 0.573 *vs*. 0.693 for Vox → Cufs, respectively. Compared with MRAA, we achieve a FID of 72.1 and an AED of 0.289 for Mixamo-Video → Fashion-Video, and 86.5 *vs*. 127.1 and 0.627 *vs*. 0.764 for Vox → Cufs, respectively. Note that, for both FID and AED metrics, smaller value is better. The large improvement indicates that the cross-domain motion transfer is challenging for the traditional FOMM and MRAA method, while our MAA model works well on the cross-domain scenario. We observe that MRAA performs worse than FOMM in the cross-domain motion transfer task, although the previous work shows MRAA usually performs better than FOMM in the traditional single-domain motion transfer[32]. This possibly dues to that the PCA based motion estimation in the MRAA method is non-parametric and less flexible for cross-domain motion transfer.

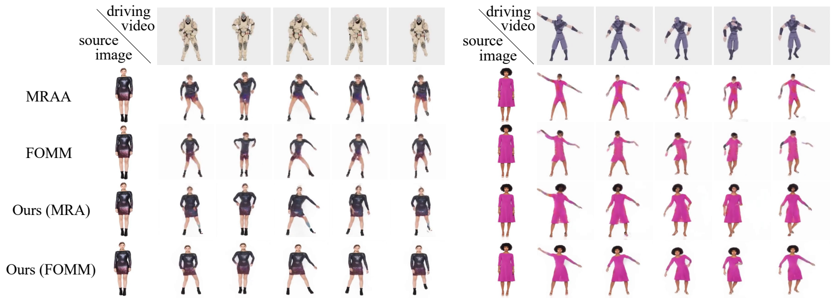

Fig. 5: Visualization results of FOMM,MRAA and ours method on human body datasets

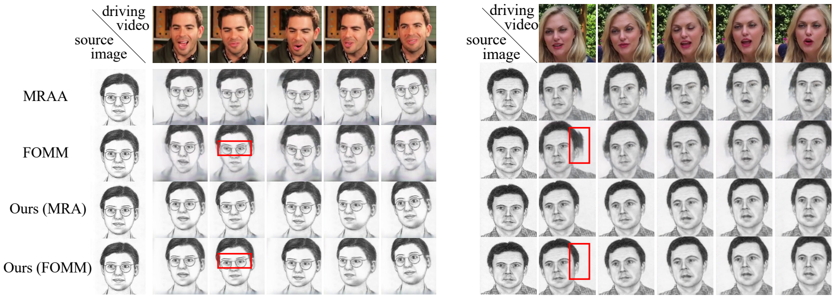

Fig. 6: Visualization results of FOMM,MRAA and ours method on human face datasets

4.3 Qualitative Results

We visualize the generated results to gain an intuitive assessment of FOMM, MRAA and our MAA models for cross-domain human body and human face animation in Fig. 5 and Fig. 6, respectively. In each figure, two pairs of results are visualized in the left and right parts, respectively. Driving frames extracted from the test video are displayed on the top row, while the source images are showed at the most left column of each part.

For human body animation, as shown in Fig. 5, the results generated by the FOMM and MRAA model obviously suffer from domain shift problem. Although the motion of driving video is roughly captured, the human body shape of source image is rarely preserved, and notable artifacts can be observed in almost all frames of the synthesized video. In contrast, our MAA model is able to capture the motion of the driving frames while properly preserving the appearance of the source image.

For human face animation, as shown in Fig. 6, the FOMM and MRAA model could generate results with a rough motion of driving frames and a similar facial appearance with source image. However, the quality of synthesized image are not satisfactory where artifacts are obvious to observe. For example, artifacts on glasses and heads can be observed for FOMM results as highlighted in the red bounding boxes. These differences in qualitative results clearly demonstrate the effectiveness of our proposed MAA model for cross-domain motion transfer.

4.4 Ablation Study

To study the impact of our proposed modules, we further conduct ablation study on both human body and human face datasets. The FOMM is used as the basic motion transfer model. The quantitative results are shown in Tab. 2, where 'w/o CYC', 'w/o SIMA' and 'w/o SGAC' means removing the cyclic training pipeline, shape-invariant motion adaptation and structure-guided appearance consistency of FOMM model, respectively.

For both human body and human face animation, as shown in Tab. 2, we observe considerable performance drops on both AED and FID for w/o CYC, which again confirms the importance to explicitly consider the cross-domain issue when performing motion transfer across domains. Similar observations can be obtained on human face

Table 2: Ablation results comparison of FOMM and our ablated models.

	Mixamo ⟶ Fashion		Vox ⟶ Cufs	
	FID ↓	AED ↓	FID ↓	AED ↓
FOMM	175.9	0.359	112.5	0.693
w/o CYC	136.9	0.354	74.1	0.633
w/o SIMA	80.2	0.303	60.7	0.622
w/o SGAC	67.7	0.284	55.2	0.603
Ours (FOMM)	**61.7**	**0.274**	**50.1**	**0.573**

dataset, which is also confirmed in the qualitative results in Fig. 7. Other ablation settings w/o SIMA and w/o SGAC also degrade the performance considerably, which validates the necessity of using the two modules for generating satisfactory synthesized video in cross-domain motion transfer.

To show the effect of each module intuitively, we further visualize the synthesized results in Fig. 7. We observe that the result of w/o CYC has richer details than that of FOMM model. For example, the face and the clothes are clearer. However, compared with our final MAA result, it still drops important motion and appearance information. Moreover, we observe the result of w/o SIMA are able to preserve relative rich appearance information, however, the pose of driving frames are not transferred properly without the help of motion consistency module. For example, artifacts can be observed for the poses of the arms and heads as highlighted in the blue rectangles. And, on the third row, w/o SGAC performs well in pose transferring but fails to preserve source image appearance without SGAC, especially for the details of human face as highlighted in red rectangles. These observations confirm the effectiveness of the modules proposed in our MAA approach.

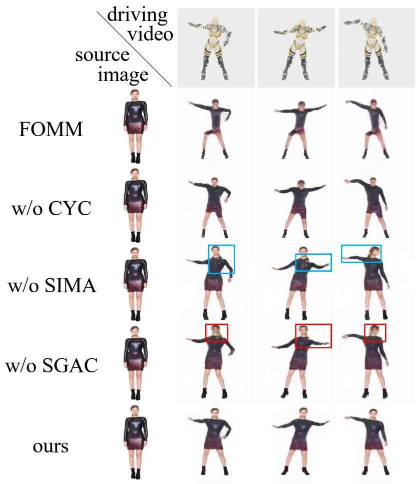

Fig. 7: Visualized ablation study results on the human body datasets

4.5 User Study

To further evaluate our model, we additionally conduct a user study. In particular, we randomly select 50 pairs of source domain driving videos and target domain source images for both human body animation and human face animation, and generate result videos in an ablation setting. For each dataset, we compare results of our final MAA model with those of FOMM and three ablation methods, respectively. The comparison are evaluated by 25 users according to three aspects, motion, appearance (denoted as app. in Tab. 3) and overall, respectively.

The user preferences are shown in Tab. 3. We observe that all scores are above 0.5, which means our results are preferred by the majority of users for all aspects in all settings. For motion aspect, fewer users prefer w/o SIMA than other settings when compared with MAA model on both datasets (0.748 vs. 0.717 and 0.679 for human body, and 0.571 vs. 0.704 for human face), which indicates SIMA improves the motion of generated results. For appearance aspect, fewer user prefer w/o SGAC than other ablation settings when compared with MAA model in human body dataset (0.715 vs. 0.711 and 0.699), which indicates SGAC contributes to appearance of generated results.

Table 3: User study results. We compare the Ours (FOMM) model to every ablation model, and the values represent the user preferences to Ours (FOMM) model

	Mixamo $\longrightarrow$ Fashion			Vox $\longrightarrow$ Cufs		
	motion	appearance	overall	motion	appearance	overall
FOMM	0.888	0.983	0.978	0.845	0.792	0.875
w/o CYC	0.717	0.699	0.732	0.571	0.615	0.626
w/o SIMA	0.748	0.711	0.702	0.704	0.655	0.675
w/o SGAC	0.679	0.715	0.725	0.593	0.617	0.575

5 Conclusion

In this paper, we propose a Motion and Appearance Adaptation (MAA) approach for cross-domain motion transfer. In MAA, we design a shape-invariant motion adaptation module to enforce the consistency of the angles of object parts in two images to capture the motion information. Meanwhile, we introduce a structure-guided appearance consistency module to regularize the similarity between the patches of the synthesized image and the source image. The experimental results demonstrates the effectiveness of our proposed method.

Acknowledgement. This work is supported by the Major Project for New Generation of AI under Grant No. 2018AAA0100400, the National Natural Science Foundation of China (Grant No. 62176047), Sichuan Science and Technology Program (No. 2021YFS0374, 2022YFS0600), Beijing Natural Science Foundation (Z190023), and Alibaba Group through Alibaba Innovation Research Program. This work is also partially supported by the Science and Technology on Electronic Information Control Laboratory.

References

1. Adobe's mixamo website, https://www.mixamo.com Accessed November 3, 2021 10
2. Amos, B., Ludwiczuk, B., Satyanarayanan, M., et al.: Openface: A general-purpose face recognition library with mobile applications. CMU School of Computer Science **6**(2) (2016) 11
3. Balakrishnan, G., Zhao, A., Dalca, A.V., Durand, F., Guttag, J.: Synthesizing images of humans in unseen poses. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8340–8348 (2018) 3
4. Burkov, E., Pasechnik, I., Grigorev, A., Lempitsky, V.: Neural head reenactment with latent pose descriptors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13786–13795 (2020) 3
5. Chan, C., Ginosar, S., Zhou, T., Efros, A.A.: Everybody dance now. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5933–5942 (2019) 1
6. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster r-cnn for object detection in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3339–3348 (2018) 4
7. Chen, Z., Wang, C., Yuan, B., Tao, D.: Puppeteergan: Arbitrary portrait animation with semantic-aware appearance transformation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13518–13527 (2020) 3
8. Chu, T., Liu, Y., Deng, J., Li, W., Duan, L.: Denoised maximum classifier discrepancy for source-free unsupervised domain adaptation. In: Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22) (2022) 4
9. Deng, J., Li, W., Chen, Y., Duan, L.: Unbiased mean teacher for cross-domain object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4091–4101 (2021) 4
10. Dong, J., Cong, Y., Sun, G., Fang, Z., Ding, Z.: Where and how to transfer: Knowledge aggregation-induced transferability perception for unsupervised domain adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–1 (2021) 4
11. Dong, J., Cong, Y., Sun, G., Zhong, B., Xu, X.: What can be transferred: Unsupervised domain adaptation for endoscopic lesions segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4022–4031 (June 2020) 4
12. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International conference on machine learning. pp. 1180–1189. PMLR (2015) 4
13. Gu, K., Zhou, Y., Huang, T.: Flnet: Landmark driven fetching and learning network for faithful talking facial animation synthesis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 10861–10868 (2020) 3
14. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017) 11
15. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017) 10
16. Huang, Z., Han, X., Xu, J., Zhang, T.: Few-shot human motion transfer by personalized geometry and texture modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2297–2306 (2021) 3
17. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017) 4
18. Jakab, T., Gupta, A., Bilen, H., Vedaldi, A.: Unsupervised learning of object landmarks through conditional image generation. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. pp. 4020–4031 (2018) 3

19. Jeon, S., Nam, S., Oh, S.W., Kim, S.J.: Cross-identity motion transfer for arbitrary objects through pose-attentive video reassembling. In: European Conference on Computer Vision. pp. 292–308. Springer (2020) 2, 3, 6
20. Kappel, M., Golyanik, V., Elgharib, M., Henningson, J.O., Seidel, H.P., Castillo, S., Theobalt, C., Magnor, M.: High-fidelity neural human motion transfer from monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1541–1550 (2021) 3
21. Li, C., Lee, G.H.: From synthetic to real: Unsupervised domain adaptation for animal pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1482–1491 (2021) 4
22. Li, Y., Huang, C., Loy, C.C.: Dense intrinsic appearance flow for human pose transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3693–3702 (2019) 3
23. Liu, W., Piao, Z., Min, J., Luo, W., Ma, L., Gao, S.: Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5904–5913 (2019) 3
24. Liu, Y., Deng, J., Gao, X., Li, W., Duan, L.: Bapa-net: Boundary adaptation and prototype alignment for cross-domain semantic segmentation. In: The IEEE International Conference on Computer Vision (ICCV) (October 2021) 4
25. Liu, Y., Deng, J., Tao, J., Chu, T., Duan, L., Li, W.: Undoing the damage of label shift for cross-domain semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition(CVPR) (2022) 4
26. Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., Van Gool, L.: Pose guided person image generation. Advances in Neural Information Processing Systems **30** (2017) 3
27. Ma, L., Sun, Q., Georgoulis, S., Van Gool, L., Schiele, B., Fritz, M.: Disentangled person image generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 99–108 (2018) 3
28. Nagrani, A., Chung, J.S., Zisserman, A.: Voxceleb: a large-scale speaker identification dataset. arXiv preprint arXiv:1706.08612 (2017) 3
29. Ren, Y., Yu, X., Chen, J., Li, T.H., Li, G.: Deep image spatial transformation for person image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7690–7699 (2020) 3
30. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: Animating arbitrary objects via deep motion transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2377–2386 (2019) 2, 3
31. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. Advances in Neural Information Processing Systems **32**, 7137–7147 (2019) 1, 2, 3, 5, 11
32. Siarohin, A., Woodford, O.J., Ren, J., Chai, M., Tulyakov, S.: Motion representations for articulated animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13653–13662 (2021) 2, 3, 5, 11
33. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014) 5
34. Tao, J., Wang, B., Xu, B., Ge, T., Jiang, Y., Li, W., Duan, L.: Structure-aware motion transfer with deformable anchor model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3637–3646 (2022) 3
35. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7167–7176 (2017) 4
36. Wang, C., Xu, C., Tao, D.: Self-supervised pose adaptation for cross-domain image animation. IEEE Transactions on Artificial Intelligence **1**(1), 34–46 (2020) 3

37. Wang, X., Tang, X.: Face photo-sketch synthesis and recognition. IEEE transactions on pattern analysis and machine intelligence **31**(11), 1955–1967 (2008) 3
38. Wiles, O., Koepke, A., Zisserman, A.: X2face: A network for controlling face generation using images, audio, and pose codes. In: Proceedings of the European conference on computer vision (ECCV). pp. 670–686 (2018) 1
39. Xu, B., Wang, B., Tao, J., Ge, T., Jiang, Y., Li, W., Duan, L.: Move as you like: Image animation in e-commerce scenario. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 2759–2761 (2021) 1
40. Zablotskaia, P., Siarohin, A., Zhao, B., Sigal, L.: Dwnet: Dense warp-based network for pose-guided human video generation. arXiv preprint arXiv:1910.09139 (2019) 3
41. Zhang, S., Zhao, W., Guan, Z., Peng, X., Peng, J.: Keypoint-graph-driven learning framework for object pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1065–1073 (2021) 4
42. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017) 2, 4, 6
43. Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., Bai, X.: Progressive pose attention transfer for person image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2347–2356 (2019) 3