Image Animation via Joint Attention Mechanism

Bing Liang, Yuhang Li, YuShan Lv
Shanghai Film Academy, Shanghai University, Shanghai, China
liangbing@shu.edu.cn, yuhangli@shu.edu.cn, 994863140@qq.com
Corresponding Author: Youdong Ding Email: ydding@shu.edu.cn

2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC) | 978-1-6654-2207-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ITAIC54216.2022.9836842

***Abstract*—Image animation refers to the automatic video synthesis task by combining the extracted source image appearance with the motion mode driving the video. Traditional image animation usually needs to predict the 3D model and then render the model. Although the non-model animation technology using deep learning has been improved in speed and effect, there are still phenomena such as unrealistic generation effects and artifacts. Therefore, this paper proposes an unsupervised image animation technology based on the fusion attention mechanism. By introducing a variety of attention mechanisms and atrous convolution, it can reduce the calculation parameters and accurately extract the motion pattern of the object to generate a real animation effect. Experiments show that our model not only achieves better visual effects, but also that our method outperforms the state-of-the-art in speed.**

Keywords—image animation; motion estimation; attention; principal component analysis

I. INTRODUCTION

Image animation is a research hotspot in computer vision field and has made amazing achievements. At the same time, image animation has been widely used in modern film and television technology industry, game industry, culture and education industry, such as reshooting and replacing actors in film and television industry, virtual dressing and makeup in e-commerce industry, etc. With the development of the short video industry, image animation can also be embedded in the short video app, so that users only need to upload a photo to follow the dance gestures of popular videos to generate their own videos. Get more users involved. At the same time, it can also be applied to the field of old photos, so that the old photos can be moved and people in memory can be "lived".

Exiting image animation technology can be divided into graphics-based methods, deep learning-based methods, and the combination of the two. The method based on graphics is to generate a three-dimensional (3D) face model first[3], and then adjust the parameters and texture mapping of the 3D model to realize the transfer of action expressions. However, the process of estimating the three-dimensional model from the 3D image and then deforming it is very time-consuming and requires high hardware. The methods based on deep learning mainly use Generative Adversarial Networks[1] and U-net[2]. Users can upload the objects and source videos they want to drive, such as animation pictures, faces, human bodies and other pose information.

However, there are still some problems in deep learning methods, such as inaccurate motion estimation for source images, blurred image generation, and unreal problems. In this paper, we propose a lightweight network framework to reduce the calculation time, and introduce attention mechanisms to improve the accuracy of motion estimation.

The main contributions of this work are:

- The channel and space fusion attention mechanism is introduced in the motion module, which makes motion estimation more accurate.
- The generation module uses a residual network that introduces attentions and atrous convolution to reduce training parameters and generate higher-quality image animations.
- Extensive experiments show that our method can achieve the state-of-the-art results or even better performance both quantitatively and qualitatively.

In this paper, the structure is organized as follows: Section II explains related research on image-based animation; Section III introduces the network structure of this paper, including the motion estimation module and generation module; Section IV shows the loss functions involved; Finally, Section V introduces the experimental results, analysis and summary are presented.

II. RELATED WORKS

A. Graphics-Based Researches

3D model technology can realize image animation by changing the 3D model of the source image. Graphics-based image technology has a long history. As early as 1999, Blanz proposed a 3D morphable model[3] (3DMM). The algorithm mainly uses Principal Component Analysis (PCA) to reduce the difference between the rendered image and the source image by adjusting the coefficients of the linear combination and rendering parameters. Vlasic adopted singular value decomposition to calculate the face shape and expression change parameters to realize the motion control of the face[4]. Gladon P proposed[5] the correspondence between parameters and motion features in the latent space. Ferrari realized the synthesis of unconstrained face images[6], adapted the

face images through 3DMM, and mapped the faces to the 3D model.

The traditional method has its advantages in image animation tasks. However, when the target video has areas that do not exist in the source video, the source human model lacks the corresponding texture after deformation, and obvious defects will appear after rendering[7]. And the process of estimating the 3D model from the 2D image and then deforming it is extremely time-consuming. Although reducing the parameters of the 3D model can improve efficiency, it will have a bad impact on the performance of the model.

B. Research Based On The Combination Of Graphics And Deep Learning

Recently, deep learning models have made significant advancements in the field of image animation. Face2Face[8] realized the real-time face playback based on monocular camera and realized real-time and high realism. Theis[9] proposed a deferred neural rendering technique, which improves the traditional UV map and renders the target image through the U-Net-based neural renderer proposed by the author. Zanfir[10] proposed an appearance transfer model for 3D human representations based on display parameterization, which complements the synthesized target image by predicting grid patches that are missing in the target image relative to the source image.

Although the introduction of deep learning technology can solve the pose problem well, these methods inherit the texture features of the source image, cannot synthesize the features that do not exist in the driving image, and require a lot of time for the rendering and texture of the 3D model.

C. Research Based On Deep Learning

Completely based on deep learning is a method to separate identity information from gesture and expression information. The lightweight model X2face[11] achieves fully self-supervised training. MonkeyNet[12] decouples morphological and dynamic information, generates a motion heat map by extracting key points driving images, and uses the motion heat map to deform source images. The author subsequently propose a first-order-motion model(FOMM)[13], which uses the first-order approximation of the local affine transformation of key points to assist in generating a dense motion field, which solves the problem that MonkeyNet cannot model the transformation of key point neighborhoods. Due to the inaccurate motion generation of FOMM for jointed objects. Therefore, the author subsequently proposed for joint animation[14], using the principal component analysis method for key areas to assist in generating a dense motion field, which can generate better animation effects for joint objects such as the human body. However, there are problems such as loss of details and unrealistic generation effects.

Although deep learning methods are able to strike a balance between efficiency and quality. the current method still has problems such as inaccurate motion estimation and unreal image generation, and it is also time-consuming.This paper proposes an improved method for the above problems.

III. METHOD

In order to solve the problem of inaccurate motion estimation, this paper introduces the channel and spatial fusion attention mechanism(SCSE) in the coding part of the motion module; for the problem of unrealistic generation effect, the residual error of the fusion attention mechanism and hole convolution is used in the image generation part. network resulting in better image animation and twice the efficiency.

A. Network Structure

The network framework is includes two parts: motion module and generation module. The purpose of the motion module is to predict the dense motion field from the frame of the driving video to the source image, which is used to calculate the motion field from the source image to the driving video, the motion field is modeled by the function $T_{S\leftarrow D}$, which maps each lost pixel in the video with the corresponding position in the source image. Instead of predicting $T_{S\leftarrow D}$ directly, motion estimation assumes a reference frame R to calculate $T_{D\leftarrow R}$ and$T_{S\leftarrow R}$. Using a self-supervised learning region detection network to obtain sparse trajectories, the motion of each key region is estimated by principal component analysis, and the two motion transformations are approximated from the sparse trajectory set. The region prediction network outputs the heatmap location of each region in the source image and driving frame and the parameters for motion estimation. Then the region transform and the source image are fed into the pixel-level flow prediction network. The resulting dense motion optical flow. In addition, the network outputs a confidence map indicating the parts of the driving video that need to be inpainted. In the image generation network, the source image is warped according to the dense optical flow and then multiplied with the confidence map to generate the target image, as shown in Fig. 1.

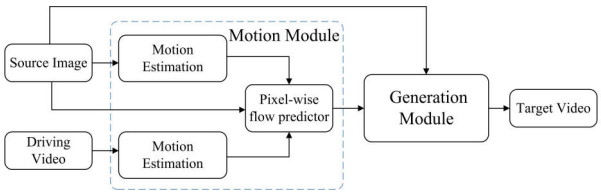

Fig. 1. Network flow.

B. Motion Module

Motion module divided into motion estimation and pixel flow prediction, and both networks adopt the U-net skip link structure. The encoder and decoder both have a five-layer structure. The encoder consists of convolution,

batch normalization, residual block and global average pooling; the decoder consists of upsampling, convolution, batch normalization, residual block and SCSE attention mechanism composition. Introduce SCSE in the decoder part to increase the weight of important feature maps or feature channels to reduce the influence of unimportant features. SCSE is a parallel connection of spatial attention (SE) and channel attention. The channel attention the core operation is a global average pooling layer and two fully connected layers to obtain the corresponding mask. The spatial attention module first compresses the space through 1×1 convolution, and then performs the sigmoid normalization operation on the feature map to complete the spatial information calibration, as shown in Fig. 2.

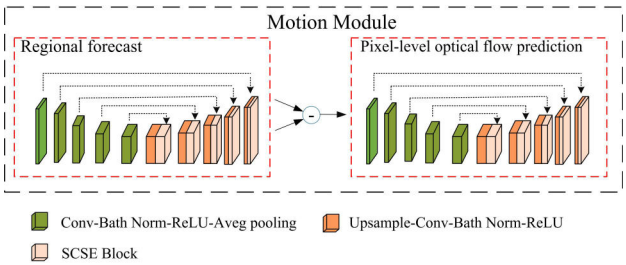

Fig. 2. The network structure of the motion module.

Accurate motion estimation is a must for high-quality image animation. In this paper, the motion is calculated from the heatmap M^k, the translation is calculated using the regression radiation parameter, and the rotation and scaling are calculated by the PCA of the heatmap M^k. Using the singular value decomposition (SVD) method to calculate PCA, equation (2) decomposes the covariance of the heatmap into unitary matrices U^k and V^k, and a diagonal matrix of singular values S^k. $\sum_{n \in N} M^k(n) = 1$, n represents a pixel location in the image, the set of all pixel locations is N, and $M^k(n)$ is the kth heatmap weight at pixel n. The formula is as follows:

$$\mu^k = \sum_{n \in N} M^k(n) n \tag{1}$$

$$U^k S^k V^k = \sum_{n \in N} M^k(n)(n - \mu^k)(n - \mu^k)^T (SVD) \tag{2}$$

C. Generation Module

Aiming at the phenomenon that the facial details are relatively blurred in the image animation generation, on the basis of the motion module used above, the SCSE attention mechanism is introduced in the image generation part, and the hole convolution residual network is combined to make the generation effect more realistic. The residual network for image generation, including two downsampling blocks, six residual fast and two upsampling modules, introduces dilated convolution in the residual block part. Upsampling and downsampling have the same structure as the motion module. As shown in Fig. 3.

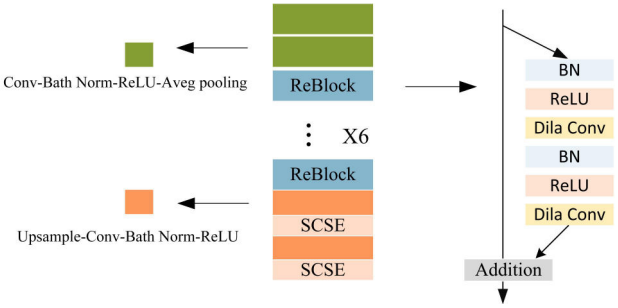

Fig. 3. Network structure for generating modules.

IV. Experiment and Result Analysis

A. Loss Function

The model is trained end-to-end using a perceptual loss based reconstruction loss as the main driving loss. Computed using a pretrained VGG-19 network. With the input driving frame D and the corresponding reconstructed frame $\widehat{D}$, the reconstruction loss is written as:

$$L_{rec}(\widehat{D}, D) = \sum_{i=1}^{I} \left| N_i(\widehat{D}) - (D) \right| \tag{3}$$

where $N_i(\widehat{D})$ is the i-th layer feature extracted from a specific VGG-19.

B. Comparison with the state of the art

Experiments evaluate our network framework on the Mgif[16]datasets. MGif is a dataset of GIF files of 2D cartoon animals collected by Google using 256*256 resolution. The experimental results are shown in Figure 4.

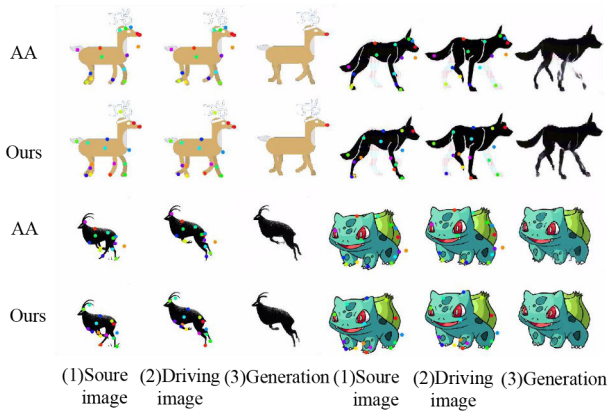

Fig. 4. Training results on the Mgif dataset.

We compared the method in this paper and AA[14] on the Mgif dataset[15]. The first and third lines are results of AA, and the second and fourth lines are the method in this paper. It can be seen intuitively that the AA method has an inaccuracy motion estimation, as shown in the first row

and the third row, the results show that key points are generated in irrelevant regions. In addition, there are problems such as generating redundant pictures or missing some areas when driving actions are generated. The method in this paper not only detects the correct key points of motion, but also can generate real and accurate driving effects.

C. Algorithm Evaluation

This paper uses the L1 loss to evaluate the algorithm. The experimental results can be seen from Table 1. The L1 evaluation index of the algorithm in this paper on the Mgif data is lower than the other algorithms for image animation. And additional experiments are also performed on Celeb-df(v2)[16], the Celeb-df(v2) dataset contains real and DeepFake synthetic videos, and the video quality is similar to the video quality of online dissemination. The Celeb-df (v2) dataset includes 590 raw videos collected from YouTube with subjects of different ages, races, and genders, and 5639 corresponding DeepFake videos. Experiments show that the algorithm in this paper can produce excellent results on both cartoon datasets and portrait datasets.

TABLE I. *EXPERIMENTAL EVALUATION VIA L1 LOSS*

Datasets	L1 loss		
	FOMM	AA	Ours
Mgif	0.0231	0.0218	0.0193
Celeb (v2)	0.0482	0.0467	0.0395

Due to the reduction of calculation parameters, the calculation speed of the algorithm in this paper has also been greatly improved. As can be seen from Table 2, the training speed of our method is improved in both stages of end-to-end. Combined with the above, in terms of quantitative evaluation, the image animation algorithm based on the spatial and channel attention mechanism proposed in this paper has better results.

TABLE II. *EXPERIMENTAL EVALUATION VIA SPEED*

Stage	speed	
	AA	Ours
1	2762.61s/it	2380.73s/it
2	894.93s/it	711.90s/it

V. CONCLUSIONS

This paper proposes an end-to-end image animation algorithm based on the joint attention mechanism of spatial channels. The algorithm introduces a joint attention mechanism in the motion module and generation module, to improves the accuracy of motion estimation, which plays a crucial role in the image animation tasks. And using the generation module based on atrous convolution can obtain better generation effect while reducing the calculation parameters. The experimental results show that this method can achieve better effect of image animation generation and reduce distortion by calculating L1 loss and other indicators. The visual effect has been improved and the speed has also been greatly improved. Our network offer a foundation for further exploration for image animation.

REFERENCES

[1] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.

[2] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.

[3] Blanz V, Vetter T. A morphable model for the synthesis of 3D faces[C]//Proceedings of the 26th annual conference on Computer graphics and interactive techniques. 1999: 187-194.

[4] Vlasic D, Brand M, Pfister H, et al. Face transfer with multilinear models[M]//ACM SIGGRAPH 2006 Courses. 2006: 24-es.

[5] Glardon P, Boulic R, Thalmann D. PCA-based walking engine using motion capture data[C]//Proceedings Computer Graphics International, 2004. IEEE, 2004: 292-298.

[6] Ferrari C, Lisanti G, Berretti S, et al. Effective 3D based frontalization for unconstrained face recognition[C]//2016 23rd International Conference on Pattern Recognition (ICPR). IEEE, 2016: 1047-1052.

[7] Averbuch-Elor H, Cohen-Or D, Kopf J, et al. Bringing portraits to life[J]. ACM Transactions on Graphics (TOG), 2017, 36(6): 1-13.

[8] Thies J, Zollhofer M, Stamminger M, et al. Face2face: Real-time face capture and reenactment of rgb videos[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2387-2395.

[9] Thies J, Zollhöfer M, Nießner M. Deferred neural rendering: Image synthesis using neural textures[J]. ACM Transactions on Graphics (TOG), 2019, 38(4): 1-12.

[10] Chan C, Ginosar S, Zhou T, et al. Everybody dance now[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 5933-5942.

[11] Wiles O, Koepke A, Zisserman A. X2face: A network for controlling face generation using images, audio, and pose codes[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 670-686.

[12] Siarohin A, Lathuilière S, Tulyakov S, et al. Animating arbitrary objects via deep motion transfer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 2377-2386.

[13] Siarohin A, Lathuilière S, Tulyakov S, et al. First order motion model for image animation[J]. Advances in Neural Information Processing Systems, 2019, 32: 7137-7147.

[14] Siarohin A, Woodford O J, Ren J, et al. Motion Representations for Articulated Animation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 13653-13662.

[15] Siarohin A, Lathuilière S, Tulyakov S, et al. Animating arbitrary objects via deep motion transfer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 2377-2386.

[16] Li Y, Yang X, Sun P, et al. Celeb-df: A large-scale challenging dataset for deepfake forensics[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 3207-3216.