applied sciences

MDPI

Article

Improved First-Order Motion Model of Image Animation with Enhanced Dense Motion and Repair Ability

Yu Xu [1], Feng Xu [1,*], Qiang Liu [1] and Jianwen Chen [2]

[1] Academy of Artificial Intelligence, Beijing Institute of Petrochemical Technology, Beijing 102617, China; 13971821877@163.com (Y.X.)
[2] School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
* Correspondence: xufeng@bipt.edu.cn

Abstract: Image animation aims to transfer the posture change of a driving video to the static object of the source image, and has potential applications in various domains, such as film and game industries. The essential part in this task is to generate a video by learning the motion from the driving video while preserving the appearance from the source image. As a result, a new object with the same motion will be generated in the animated video. However, it is a significant challenge if the object pose shows large-scale change. Even the most recent method failed to achieve this correctly with good visual effects. In order to solve the problem of poor visual effects in the videos with the large-scale pose change, a novel method based on an improved first-order motion model (FOMM) with enhanced dense motion and repair ability was proposed in this paper. Firstly, when generating optical flow, we propose an attention mechanism that optimizes the feature representation of the image in both channel and spatial domains through maximum pooling. This enables better distortion of the source image into the feature domain of the driving image. Secondly, we further propose a multi-scale occlusion restoration module that generates a multi-resolution occlusion map by upsampling the low-resolution occlusion map. Following this, the generator redraws the occluded part of the reconstruction result across multiple scales through the multi-resolution occlusion map to achieve more accurate and vivid visual effects. In addition, the proposed model can be trained effectively in an unsupervised manner. We evaluated the proposed model on three benchmark datasets. The experimental results showed that multiple evaluation indicators were improved by our proposed method, and the visual effect of the animated videos obviously outperformed the FOMM. On the Voxceleb1 dataset, the pixel error, average keypoints distance and average Euclidean distance by our proposed method were reduced by 6.5%, 5.1% and 0.7%, respectively. On the TaiChiHD dataset, the pixel error, average keypoints distance and missing keypoints rate measured by our proposed method were reduced by 4.9%, 13.5% and 25.8%, respectively.

Keywords: image animation; convolutional block attention module; first order motion model; generative adversarial networks

Citation: Xu, Y.; Xu, F.; Liu, Q.; Chen, J. Improved First-Order Motion Model of Image Animation with Enhanced Dense Motion and Repair Ability. *Appl. Sci.* **2023**, *13*, 4137. https://doi.org/10.3390/app13074137

Academic Editor: Yu-Dong Zhang

Received: 21 February 2023
Revised: 15 March 2023
Accepted: 21 March 2023
Published: 24 March 2023

1. Introduction

Image animation is the technology that transfers the motion posture of an object in a driving video to the static object in the source frame. Given a source image and a driving video depicting the same object type, the goal of image animation is to generate a video by learning the motion from the driving video while preserving the appearance from the source image. As shown in Figure 1, the motion from the driving videos in the second row is transferred to the source images in the first row. In the animated videos in the third row, the objects from the source images follow the exact same motion as the driving videos. Nowadays, image animation has achieved extensive application in film production, virtual reality, photography and electronic commerce.

Figure 1. Our method animates still source images.

Traditional approaches for image animation typically involve using data fusion from different data sources to acquire prior knowledge of the object [1], (such as a 3D model), and using computer graphics technology to solve the problem [2–4]. Face2Face [4] uses a 3D parametric model with 269 parameters to fit facial posture, expression, illumination and shape information. By detecting and tracking the face image and adjusting the relevant parameters, the expression of the face in the source domain is converted into the expression of the face in the target domain. However, this approach is not applicable when the task is not limited to the face. With the development of machine learning and deep learning, many industries have achieved substantial development [5–7].

Recently, with the development of generative adversarial networks [8] (GANs) and variational auto-encoders [9] (VAEs), many methods have been proposed and have expedited the research and application of image animation. These methods have been used as substitutes for 3D parametric models in generating real images. These methods are mainly divided into two categories. The first type of method relies on pre-trained models to extract specific representations of the target object, such as facial landmarks [10,11], gesture [12] or human keypoints [13]. However, the performance of these methods depends on the labeled data and pre-trained model. The construction of the pre-trained model requires the annotation of the ground-truth data, which requires expensive acquisition. In addition, these pre-trained models do not generally apply to all types of object categories. Another method is unsupervised motion transfer, which does not require real data on the ground. X2Face [14], proposed by Oliva Wiles et al., is a self-supervised neural network model that uses another person's face to control the pose and expression of a given face, but the error generated by this method is obvious. Aliaksandr Siarohin et al. proposed Monkey-Net [15], which is the first depth model for image animation of unknowable objects. It extracts target keypoints in the image through a self-supervised keypoint detector and generates a dense heatmap from sparse keypoints. Following this, the input frame, which uses the motion heatmap and appearance information extracted from the input image, is synthesized. However, it is difficult for Monkey-Net to model the appearance transformation of objects near the keypoints, which leads to poor generation quality when the scale of object change is quite large. To support more complex motion, FOMM [16] was proposed to use unsupervised learning keypoints and local affine transformation to simulate complex motion.

In applications, the methods of image animation based on pre-trained models have many limitations. Meanwhile, research on the application of image animation without relying on labeled data and pre-trained models has achieved great progress. However, there are still some problems in the current unsupervised methods. For example, sometimes

the prediction of the optical flow field is not accurate, which will result in incorrect or low-quality generated frames. Sometimes, the predicted keypoints are located on the background instead of on the moving objects. Hence, the displacement deformation between keypoints cannot accurately describe the displacement deformation of the rigid region of the animation object. The ghosting effect (false object shadow) often occurs in the generated frames.

Although FOMM is an advanced model in the field of image animation, it still fails to accurately transfer motion information in videos with significant posture changes of human bodies or faces. For instance, when given a source image with a frontal human body and a driving frame with a side or back human body, FOMM cannot accurately generate frames with the correct posture, as shown in Figure 2a,b. Additionally, FOMM often generates images with ghost effects (false object shadow), as depicted in Figure 2c,d. Furthermore, as illustrated in Figure 3, some keypoints of the target object predicted by FOMM are located on the background, which potentially causes the incorrect parts in the generated frames.

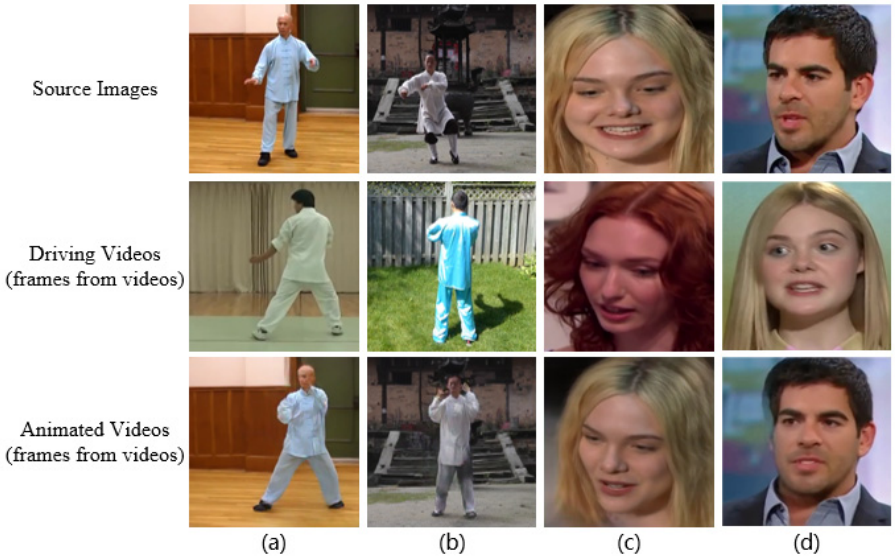

Figure 2. Failure cases of FOMM. (**a**) Frame 1 generated with the incorrect posture. (**b**) Frame 2 generated with the incorrect posture. (**c**)Frame 1 generated with ghost effects. (**d**) Frame 2 generated with ghost effects.

In order to solve these problems and improve the performance, we propose an enhanced model based on FOMM. In the proposed novel framework, we focus on two main contributions.

(1) We propose an attention module to optimize the generation of optical flow field, which could improve the precision of the generated optical flow field and obtain a more stable and robust motion representation.
(2) We propose a multi-occlusion network to repair the details of the picture from multiple scales, and to obtain more accurate results. We take advantage of the occlusion map's ability to correct the pixel values in the generated images and achieve better visual results.

Generally, the advantages of the proposed method can be concluded as following.

(1) Using the proposed method, it has been proven by experiments that the keypoints output by the keypoint detector module are correctly located in the region of the target object after using the multi-scale occlusion restoration module.
(2) With the objective evaluation, the proposed method outperformed FOMM on the Voxceleb1 dataset, with a reduction of 6.5%, 5.1% and 0.7% in pixel error, average keypoint distance, and average Euclidean distance, respectively. The results on the

TaiChiHD dataset also showed significant improvement, with a reduction of 4.9% in pixel error, 13.5% in average keypoint distance, and 25.8% in missing keypoint rate.

(3) With the subjective evaluation, the proposed method also demonstrated superior performance compared to FOMM, indicating its potential for a range of image animation. These results prove that our framework has more effective repair capabilities and generates images with better visual effects.

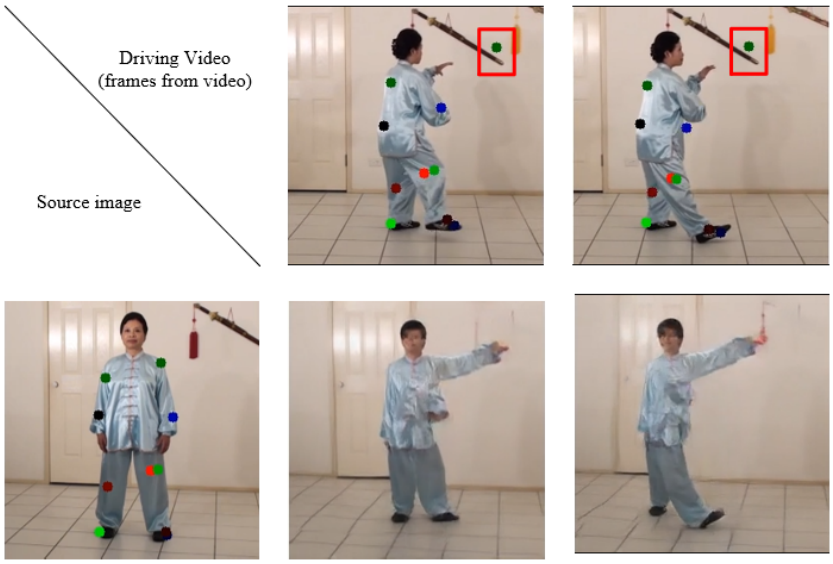

Figure 3. Incorrectly located key points in FOMM, resulting in poor frame generation results. (Different colorful dots represent different key points located by FOMM, and the red boxes indicate the incorrect key points.)

The remainder of this paper is organized as follows, in Section 2, the proposed model and algorithm are described in detail; in Section 3, the experimental setup and results are described, and the experimental results are analyzed; finally, in Section 4, we conclude the paper, and then discuss the next research directions and provide a reasonable prospect of our study.

2. Methodology

FOMM is an end-to-end network, which does not require a priori knowledge of the dataset. When the data are used for training, the entire network can be directly applied to the same types of datasets. However, in the case of large-scale posture change, such as excessive change in facial expressions or the overall rotation of the human body by 180 degrees, the video frames generated by FOMM are of low quality, and even contain incorrect frame content. In a study of FOMM algorithm, we found that the reasons for failure include the keypoint positioning error of FOMM and the inaccurate estimation of the optical flow field. In order to solve these problems, we propose an attention module to extract the strong information of intermediate features, so that more accurate features can be used to predict the optical flow field. Furthermore, in order to enhance the repair ability of the network, we propose a multi-scale occlusion restoration module to optimize the quality of reconstructed images.

2.1. Algorithm Framework

Our improved FOMM is mainly composed of the keypoint detector module, the dense motion module, the attention module, the multi-scale occlusion restoration module and the generator module. The network structure is shown in Figure 4. The keypoint detector module and the dense motion module are basically the same as the corresponding modules in FOMM.

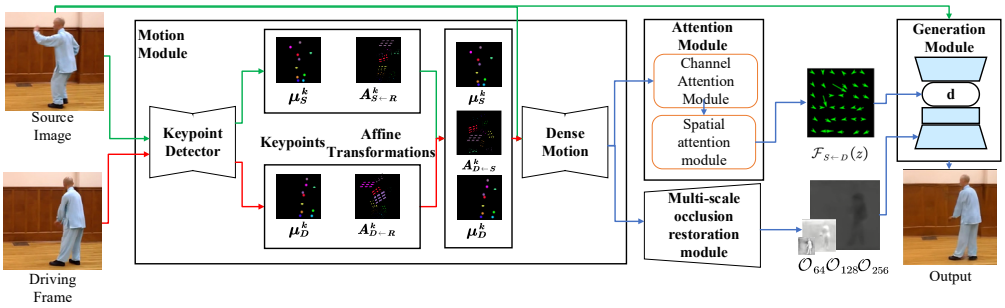

Figure 4. The framework of our method.

We denote S and D as the source and the driving frames extracted from the same video, respectively. The keypoint detection module detects the unsupervised key points of the image. The dense motion module first establishes sparse motion and affine transformation between individual rigid regions through the key points, and then generates dense optical flow and an occlusion map of the whole image through the S warped by the sparse motion.

Keypoints prediction based on U-net output K heat maps, $M^1, \ldots, M^K$ for the input image at unsupervised keypoints, followed by softmax, s.t. $M^k \in [0,1]^{H \times W}$, where H and W are the height and width of the image, respectively, and $\sum_{z \in Z} M^k(z) = 1$, where z is a pixel location (x, y coordinates) in the image, the set of all pixel locations being Z, and $M^k(z)$ is the k-th heatmap weight at pixel z. Equation (1) estimates the translation component of affine transformation in the abstract coordinates mapped by the input picture.

$$\mu^k = \sum_{z \in Z} M^k(z) z. \tag{1}$$

For both frames S and D, the keypoints prediction also output four additional channels $P_{ij}^k \in \mathbb{R}^{H \times W}$ for each keypoint, where $i \in \{0,1\}, j \in \{0,1\}$ indexes the affine matrix $A_{X \leftarrow R}^k[i,j]$, where $X \in \{S, D\}$, and R is the assumed reference frame. This is shown in Equation (2).

$$A_{X \leftarrow R}^k[i,j] = \sum_{z \in Z} M^k(z) P_{ij}^k(z). \tag{2}$$

The dense motion module uses an encoder–decoder structure to predict the rigid mask of each keypoint through the heatmap representation of the keypoints and the warped S frame of the sparse optical flow field. Equations (3) and (4) represent the heatmap representation of the keypoints of the input image to R and the heatmap representation of S to D, respectively. Where σ is a hyper-parameter, we usually take 0.01 based on experience.

$$G_X^k = e^{-0.5 \times \sum_{z \in Z} (z - \mu_X^k)^2 / \sigma} \tag{3}$$

$$H = [(G_D - G_S),\ 0] \tag{4}$$

In order to obtain the sparse optical flow field, FOMM needs to obtain the affine matrix $A_{S \leftarrow D}^k$ (Equation (5)) from S to D, and calculate the sparse optical flow field through Equation (6).

$$A_{S \leftarrow D}^k = A_{S \leftarrow R}^k \left(A_{D \leftarrow R}^k \right)^{-1} \tag{5}$$

$$\mathcal{F}^k = \mu_S^k + A_{S \leftarrow D}^k \left(z - \mu_D^k \right) \tag{6}$$

With inputting S (warped by $\mathcal{F}^k$) and H, the dense motion module outputs an intermediate feature ξ. In Equation (7), $\mathcal{F}$ represents warping operation and *Unet* represents model architecture.

$$\xi = Unet(H, f_w(S, \mathcal{F})) \tag{7}$$

Then, the intermediate features ξ will enter the attention module to obtain the dense optical flow field and the multi-scale occlusion restoration module to obtain multiple occlusion maps. We will introduce our attention module and multi-scale occlusion restoration module in detail in Sections 2.2 and 2.3.

2.2. Attention Module

Based on the basic CBAM (convolutional block attention module) [17], we propose an attention module that is more suitable for our model. Similarly, the attention module also infers the attention map of the middle feature layer from the channel and space dimensions, and is used for adaptive feature refinement.

The attention module has two sequential sub-modules: the channel attention module and the spatial attention module. The process is shown in Figure 5.

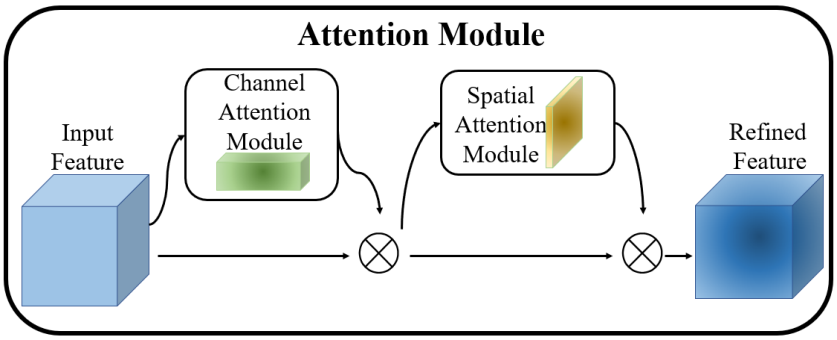

Figure 5. The overview of attention module.

The channel attention module leverages channel-wise max-pooling to identify the most salient activation for each channel in the feature map. This information is then fed through a multi-layer perceptron (MLP) [18] to get the weight of the max-pooled feature. These attention weights are then multiplied by a higher-parameter σ. Usually, we give σ the value 2. Finally, the channel attention weight is obtained through sigmoid and the channel-refined feature is obtained by weighting the channel attention weight to the input feature, as shown in Figure 6. The channel attention is computed as Equation (8).

$$W_C = sigmoid(\sigma(MLP(MaxPool(\xi)))) \tag{8}$$

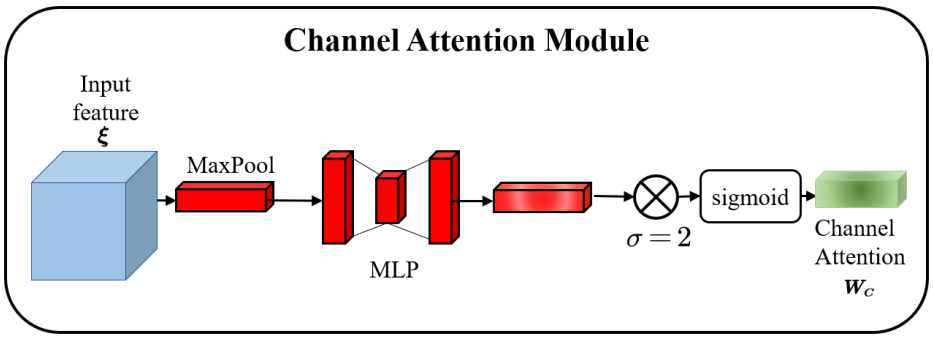

Figure 6. Diagram of the channel attention module.

In the design of the spatial attention module, we follow the design of CBAM. We use the spatial max pooling and the spatial average pooling to obtain the maximum and mean values in the channel-refined feature after the channel attention module, and then obtain the spatial attention weight after the convolution and sigmoid operation. Finally, we apply the spatial attention weight onto the channel-refined feature to obtain the refined feature,

as shown in Figure 7. The channel attention is computed as Equation (9), where $Conv_{7\times7}$ represents a convolution operation with the filter size of 7×7.

$$W_S = sigmoid\left(Conv_{7\times7}\left(\left[AvgPool\left(\xi'\right); MaxPool\left(\xi'\right)\right]\right)\right) \tag{9}$$

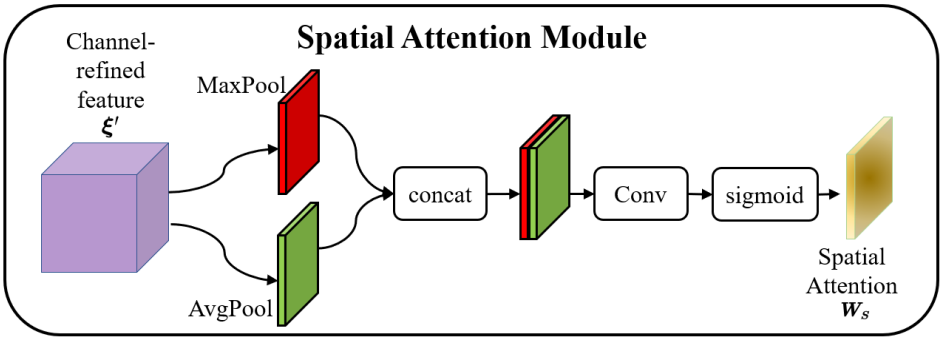

Figure 7. Diagram of spatial attention module.

We input the intermediate feature ξ from the dense motion module into the attention module to obtain the refined feature ξ'. The calculation formula is shown in Equation (10)

$$\xi' = Attention(\xi) = W_S \otimes W_C \otimes \xi \tag{10}$$

Then, we predict the masks, $Mask^0 \ldots Mask^k$ corresponding to the key points by using ξ', as shown in Equation (11).

$$Mask = Conv_{7\times7}\left(\xi'\right) \tag{11}$$

The final dense motion prediction $\mathcal{F}_{S\leftarrow D}(z)$ is given by:

$$\mathcal{F}_{S\leftarrow D}(z) = Mask^0 z + \sum_{k=1}^{K} Mask_k \odot \mathcal{F}^k \tag{12}$$

Note that the term $Mask^0 z$ is considered in order to model non-moving parts, such as the background.

2.3. Multi-Scale Occlusion Restoration Module

In order to improve the low-quality visual effect of reconstructed images when the poses of animated objects change greatly, we propose a multi-scale occlusion restoration module. The structure is shown in Figure 8. The multi-scale occlusion restoration module takes the intermediate feature ξ of the dense motion module as input. The feature ξ will pass through two upblock2d modules. Each upblock2d block is composed of bilinear interpolation, a convolution module, BatchNorm and a ReLU activation function. After the feature ξ passes through two upblock2d modules, the two intermediate features with dimensions of (64, 128, 128) and (32, 256, 256) are output, respectively. Finally, convolution operation with the filter size of 1×1 is used to reduce the channels of each feature layer to 1, so as to obtain the occlusion map with three resolutions, with the resolutions being (64, 64), (128, 128), (256, 256).

The function of the multi-scale occlusion restoration module is to gradually repair the details of the reconstructed image from multiple resolutions, so as to make the details of the reconstructed image more vivid and natural. The multi-scale occlusion map output by the multi-scale occlusion restoration module will then be input into the generator to guide the generation of the reconstructed image.

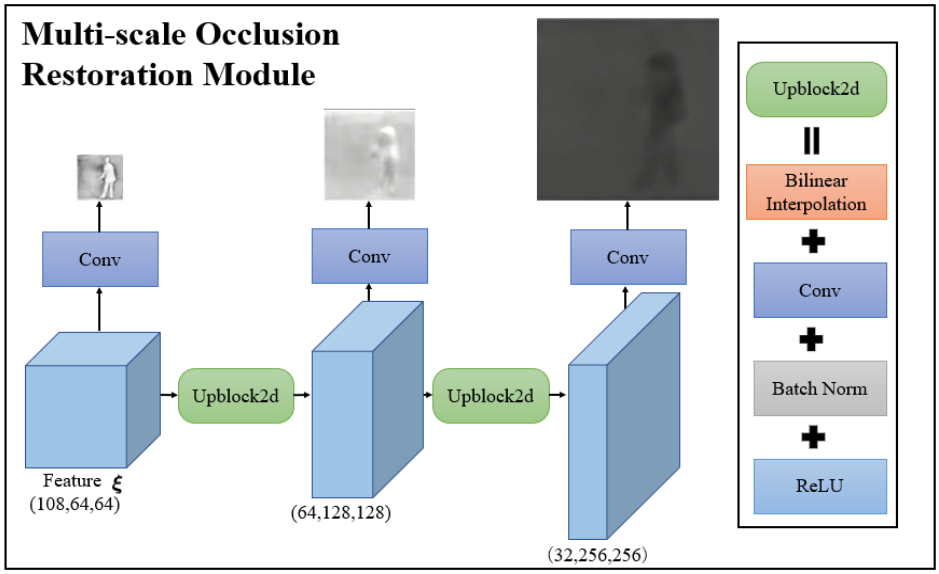

Figure 8. Diagram of the multi-scale occlusion restoration module.

2.4. Generator

The construction of the generator follows the automatic encoder structure, as shown in Figure 9. The generator is composed of an encoder and a decoder. The encoder has two blocks called downblocks. The downblocks will double the feature on the channel, halve the height and width of the feature and finally output an intermediate feature with a resolution of (256, 64, 64). At this point, we use the dense motion $\mathcal{F}_{S \leftarrow D}(z)$ distort the intermediate feature, so as to transfer the motion information from the source frame to the driving frame. In addition, in order to encode intermediate feature information at a deeper level, we use six modules called Resblock2d. In 2016, Kaiming He et al. [19] proposed the use of Resblock2d in ResNet to solve the problem of network gradient disappearance and gradient explosion, so that deeper network models can be trained. The structure diagram of Resblock2d is shown in Figure 9.

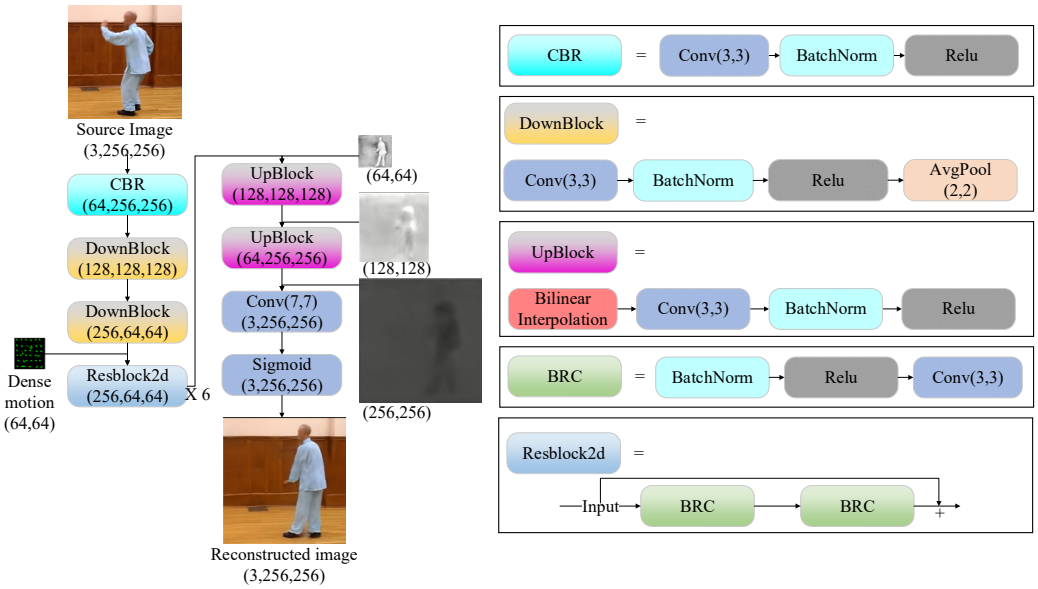

Figure 9. Diagram of the generator.

In the decoding stage, we first use the occlusion map of (64, 64) resolution to repair the intermediate feature. Specifically, the calculation is shown in Equation (13), where ξ_{in}

represents the input feature, ξ_{out} represents the feature after the occlusion map repair, and $\mathcal{O}_i$ represents the occlusion map with resolution (i, i) (i belongs to (64, 128, 256)). After that, the repaired feature will go through two blocks called upblocks. The upblocks will halve the channels of feature and double the height and width, which is the reverse operation of the downblocks. The output of each upblock will be repaired in detail through an occlusion map of the same resolution. Finally, convolution operation with the filter size of 1×1 is used to reduce the channels of feature to 3, and our reconstructed image is obtained after the sigmoid activation function.

$$\xi_{out} = \mathcal{O}_i \odot \xi_{in} \tag{13}$$

3. Experiments and Results

In order to evaluate the performance of our proposed method and compare with FOMM, we conducted extensive experiments on the three benchmark datasets, including VoxCeleb1 [20], TaiChiHD [16] and MGif [14]. Each dataset has a separate training set and test set. Some sample images (frames of videos) in the three datasets are shown in Figure 10.

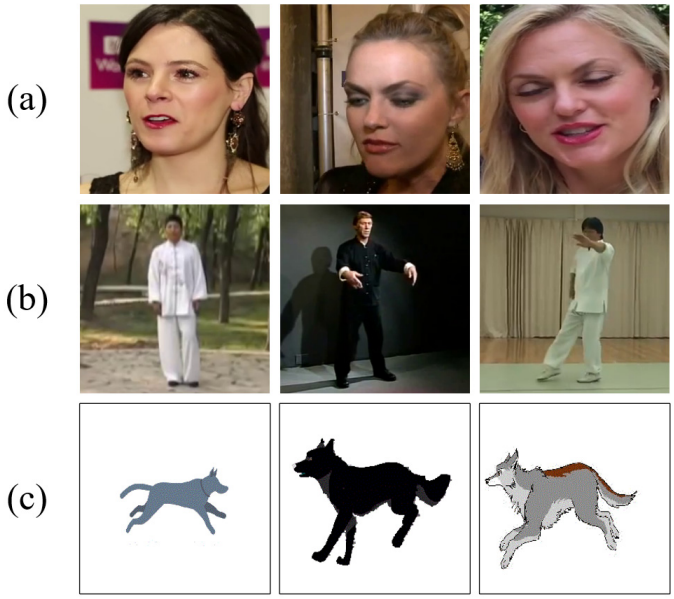

Figure 10. Examples of experimental datasets. (**a**) TaiChiHD; (**b**) VoxCeleb; (**c**) MGif.

3.1. Datasets and Experimetal Setting

VoxCeleb1 is an open-source large-scale celebrity interview voice collection, collected by Nagrani and others from Google. In order to obtain the corresponding video files, the same processing method proposed in FOMM was used to download the celebrity interview videos of VoxCeleb1. For each video, the face area was extracted and marked with a square area, and then was normalized to size 256×256. The frame number range of each video was 64-1024. There were 18,130 training videos and 503 test videos in total. In our experiment, human faces were generated and animated for test videos.

TaiChiHD is a dataset composed of cut videos of human bodies performing Tai Chi movements, published by Aliaksandr Siarohin et al. We also used the above processing method to obtain video images in size 256×256, including 2652 training videos and 285 test videos. In our experiment, human bodies with Tai Chi movements were generated and animated for test videos.

MGif is a GIF (graphics interchange format) file dataset that describes 2D cartoon animals. The dataset was collected through Google search, including 900 training videos and 100 test videos. In our experiment, cartoon animals were generated and animated for test videos.

3.2. Measurement Metrics

Generally, animation image quality assessment includes reconstruction quality and animation quality. In terms of reconstruction quality, given that image animation is a relatively new research problem, there are not many effective ways to evaluate this currently. For quantitative metrics, video reconstruction accuracy was used as a proxy for image animation quality. We applied the same metrics in our experiments.

We defined $\mathcal{L}_1$ error as the mean absolute difference between reconstructed and ground-truth video pixel values. As shown in Equation (14), n represents the total number of video frames, while H and W represent the height and width of the image, respectively. I_{hw} represents the pixel value at (h, w) position in the real video frame, and $\hat{I}_{hw}$ represents the pixel value at (h, w) position in the reconstructed frame.

$$\mathcal{L}_1 = \left(\sum_{i=1}^{n} \sum_{h=0}^{H} \sum_{w=0}^{W} \left| I_{hw} - \hat{I}_{hw} \right| \right) / n \tag{14}$$

Average keypoint distance (AKD) and missing keypoint rate (MKR) were used to evaluate the difference between poses of reconstructed and ground truth videos. Landmarks were extracted from both videos using public, body [21] (for TaiChiHD) and face [22] (for VoxCeleb) detectors. AKD is the average distance between corresponding landmarks, while MKR is the proportion of landmarks existing in the ground-truth video but missing in the reconstructed video.

Average Euclidean distance (AED) was used to evaluate how well identity is preserved in reconstructed videos. Public reidentification networks for bodies [23] (for TaiChiHD) and for faces [24] (for VoxCeleb) extracted identity from reconstructed and ground-truth frame pairs. Then, the mean $\mathcal{L}_2$ norm of their difference across all pairs was computed.

Animation quality is usually evaluated by subjective video quality assessment. We will present some examples of image animation generation results for visual comparison.

3.3. Experimental Results and Analysis

3.3.1. Reconstruction Quality

Quantitative reconstruction results are shown in Table 1. From the results in Table 1, we can see that our method achieved better results than FOMM with almost all indicators on all the three datasets. Especially on the TaiChiHD dataset, $\mathcal{L}_1$, AKD and MKR decreased by 4.9%, 13.5% and 25.8%, respectively. In addition, on VoxCeleb1 for face movement transfer, $\mathcal{L}_1$, AKD and AED decreased by 6.5%, 5.1% and 0.7%, respectively.

Table 1. Comparing the video reconstruction evaluation indicators. FOMM and FOMM-CBAM (ours) set the number of key points to 10 (the best result is displayed in bold).

	TaiChiHD			**VoxCeleb**			**MGif**
	$\mathcal{L}_1$	**(AKD, MKR)**	**AED**	$\mathcal{L}_1$	**AKD**	**AED**	$\mathcal{L}_1$
X2Face	0.080	(17.65, 0.109)	0.27	0.078	7.69	0.405	-
FOMM	0.061	(6.75, 0.031)	0.167	0.046	1.37	0.142	0.026
Ours	**0.058**	**(5.84, 0.023)**	0.170	**0.043**	**1.30**	**0.141**	**0.026**

3.3.2. Animation Quality

In order to compare our method with FOMM in the terms of animation quality, we performed animation generation on the TaiChiHD dataset and the VoxCeleb1 dataset. The experimental results are shown in Figures 11 and 12, respectively. The results show that the animation quality significantly improved in most cases, especially in the case of animated objects with large-scale posture change.

Figure 11. Comparison between our method and FOMM on the TaiChiHD dataset.

Figure 12. Comparison between our method and FOMM on the VoxCeleb1 dataset.

In Figure 11, we compared the performance of our model and FOMM on the TaiChiHD dataset. The generated images in the first row show that our model produced more vivid and natural results than FOMM when the posture transformation was not large enough. In the second row, when the human body turned around, FOMM failed to locate the posture change of the human body, particularly the change in head posture, which resulted in completely incorrect human body posture in the reconstructed frame. In general, our model can transfer more complex motion postures by using the attention mechanism to extract more effective features and the multi-scale occlusion map provided by the multi-scale occlusion restoration module reconstruction frame. Hence, the generated results of our model were much better than those of FOMM. In the third row of Figure 11, we observed that the video frame generated by FOMM had a large shadow, whereas our model alleviated this problem.

In Figure 12, the comparative results on the VoxCeleb1 dataset between our model and FOMM are provided. The images in the first and second rows show the generated video frames when the face turned to an extreme angle, such as the side face turning to the front or the front face turning to the side. FOMM lost the characteristics of the source image object, resulting in obvious errors in the generated frame. Our method used the attention mechanism to optimize the generation of the optical flow field, improving the

accuracy of the generated optical flow field and making the motion of the generated face more consistent to the source frame. Furthermore, as shown in the third row of Figure 12, FOMM resulted in a ghosting effect (false object shadow) in the generated face, whereas our method solved this problem by gradually repairing the generated results with the multi-scale solution.

In general, the experimental results demonstrated that our model is capable of generating the optical flow fields with higher accuracy. Furthermore, our model effectively resolved the ghosting effect (false object shadow) in the generated frames. As a result, our model achieved a superior image animation generation effect compared to previous methods.

3.3.3. Ablation Experiment

In order to further analyze the benefits with the attention module and the multi-scale occlusion restoration module in our method, we conducted a large quantity of ablation experiments. The results are shown in Table 2.

Table 2. Ablation study on the TaiChiHD (256) dataset with 10 keypoints (the best result is displayed in bold).

Methods	$\mathcal{L}_1$	TaiChiHD (AKD, MKR)	AED
FOMM	0.061	(6.75, 0.031)	0.167
FOMM + CBAM	0.059	(6.16, 0.024)	0.178
FOMM + our attention module	0.060	(6.60, 0.029)	0.181
FOMM + CBAM + multi-scale occlusion restoration module	0.059	(6.54, 0.027)	(0.179)
Our method	**0.058**	**(5.84, 0.023)**	0.170

From Table 2, we can see that when we used the attention module exclusively, all of our metrics were worse than those by using CBAM together. However, when we included the multi-scale occlusion restoration module again, our method achieved much better performance. Compared with the FOMM and CBAM methods, all the metrics indicate that our method had the better performance. The proposed attention module seeks to identify the most dominant features in a frame by eliminating the average pooling in space and solely relying on maximum pooling. The task of reconstructing the background and addressing occlusions is transferred to the multi-scale occlusion restoration module, which employs a multi-scale occlusion map to perform step-by-step repair. Thus, there is no conflict between the two modules.

In addition, we also visualized the keypoints, as shown in Figure 13. From the comparison of experimental results, it is obvious that the keypoints detected by our method were more consistent with the human body structure, such as defining a keypoint for the head. However, in FOMM, some keypoints were incorrect. FOMM also had the problem of keypoints positioning error, such as the keypoints on the background, which will lead to the leakage of motion information to the background.

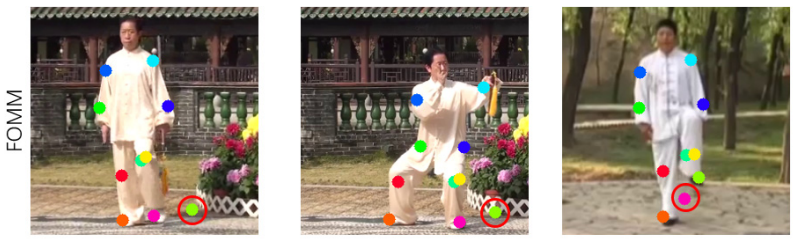

Figure 13. *Cont.*

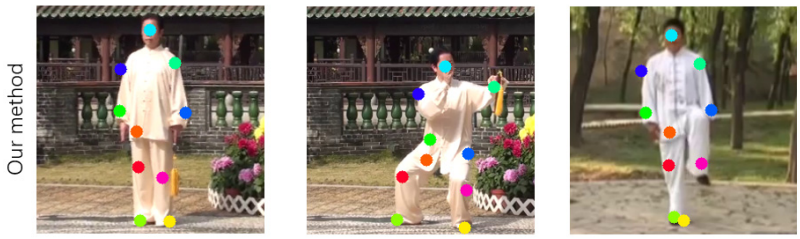

Figure 13. Our method and FOMM's visualization of keypoints in the TaiChiHD dataset. (The different colorful dots represent different key points, and the red circles indicate the incorrect key points.)

3.3.4. Analysis of Reconstruction Results

Although our method has made significant improvements in both the reconstruction quality and the animation quality compared with FOMM, the proposed method still does not achieve perfect visual effects in some special cases. A few examples with poor visual effects in the reconstruction process are shown in Figure 14. After investigating those cases, we found that our method cannot achieve good visual effects in two main situations. Firstly, our method could not handle the reconstruction task when the character's clothing color was very similar to the background color, as shown in Figure 14a. In this case, the color information of the target character in the reconstructed frame mixed with the color information of the background. Secondly, our method could not handle the situation when the source frame was lacking some information. As shown in Figure 14b, the target character in the source frame stands on the side and the facial information is occluded by one of arms, so the facial information was deficient and could not be accurately generated in the reconstructed frame. In these types of special cases, our method still faced great challenges in the TaiChiHD dataset.

Figure 14. The reconstruction failure case of our method. (**a**) An error caused by the character's clothing color was very similar to the background color. (**b**) An error caused by the source frame was lacking some information.

4. Conclusions and Future Work

In this paper, an improved framework for image animation generation has been proposed based on the FOMM method. Specifically, the two novel modules have been proposed and applied to solve the problems of inaccurate reconstruction and low quality of visual effects in video frame generation.

Firstly, we proposed to use an attention module to optimize the generation of the optical flow field. The attention module can further enhance feature expression by reconstructing the feature information on the channel and space, so as to predict the more precise

optical flow field. Secondly, we proposed the multi-scale occlusion restoration module to obtain an occlusion map, with resolutions of (64, 64), (128, 128) and (256, 256), to repair the feature representation of the network at different resolutions and to enhance the repair ability of the network. With this proposed module, the generated frames can contain the correct and complete visual information and be of better visual quality in the case of large posture changes of the animated object. In addition, our model can be trained effectively in an unsupervised manner. Based on the above two modules, we proposed our improved framework. In order to verify the performance of our method, we conducted extensive experiments on three benchmark datasets, TaiChiHD, VoxCeleb1 and MGif. The experimental results showed that our method outperformed the FOMM in both the reconstruction quality and the animation quality.

Although our proposed framework has achieved apparent improvement for image animation, there are still some limitations. As one of the limitations, the inter-frame correlation was not considered. In future work, we plan to utilize neural networks, such as LSTM [25], to save the generation results of the previous frame, and then the saved information can be used enhance the generation of the current frame. Given the correlation between the two consecutive frames, some generated content from the previous frame will help to improve the reconstruction quality of the current frame. Additionally, we also plan to explore the use of multiple source frame images from various angles to build potential source frames. We will automatically adjust the contribution of source frames from different angles through neural networks to generate more accurate and realistic reconstructed frames in our future work.

Author Contributions: Conceptualization, F.X.; methodology, Y.X., J.C. and F.X.; software Y.X.; validation, F.X.; formal analysis, Y.X.; investigation, J.C. and F.X.; resources, F.X.; data curation, Y.X; writing—original draft preparation, Y.X.; writing—review and editing, F.X.; supervision, Q.L.; project administration, F.X.; funding acquisition, Q.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the fund of the Beijing Municipal Education Commission, China, under grant number 22019821001.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in https://github.com/AliaksandrSiarohin/first-order-model, accessed on 9 December 2021.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. AlZu'bi, S.; Jararweh, Y. Data fusion in autonomous vehicles research, literature tracing from imaginary idea to smart surrounding community. In Proceedings of the 2020 Fifth International Conference on Fog and Mobile Edge Computing (FMEC), Paris, France, 20–23 April 2020; pp. 306–311. [CrossRef]
2. Cao, C.; Hou, Q.; Zhou, K. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Trans. Graph.* **2014**, *33*, 1–10. [CrossRef]
3. Blanz, V.; Vetter, T. A morphable model for the synthesis of 3d faces. In Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, Los Angeles, CA, USA, 8–13 August 1999; pp. 187–194. [CrossRef]
4. Thies, J.; Zollhöfer, M.; Stamminger, M.; Theobalt, C.; Nießner, M. Face2face: Real-time face capture and reenactment of rgb videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 July 2016; pp. 2387–2395. [CrossRef]
5. Rezaee, H.; Aghagolzadeh, A.; Seyedarabi, M.H.; AlZu'bi, S. Tracking and occlusion handling in multi-sensor networks by particle filter. In Proceedings of the 2011 IEEE GCC Conference and Exhibition (GCC), Dubai, United Arab Emirates, 19–22 February 2011; pp. 397–400. [CrossRef]
6. Elbes, M.; Almaita, E.; Alrawashdeh, T.; Kanan, T.; AlZu'bi, S.; Hawashin, B. An indoor localization approach based on deep learning for indoor location-based services. In Proceedings of the 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), Amman, Jordan, 9–11 April 2019; pp. 437–441. [CrossRef]

7. AlZu'bi, S.; Aqel, D.; Mughaid, A. Recent intelligent approaches for managing and optimizing smart blood donation process. In Proceedings of the 2021 International Conference on Information Technology (ICIT), Amman, Jordan, 14–15 July 2021; pp. 679–684. [CrossRef]
8. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
9. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2014**, arXiv:1312.6114.
10. Zhang, J.; Zeng, X.; Wang, M.; Pan, Y.; Liu, L.; Liu, Y.; Ding, Y.; Fan, C. Freenet: Multi-identity face reenactment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5326–5335. [CrossRef]
11. Huang, P.-H.; Yang, F.-E.; Wang, Y.-C.F. Learning identity-invariant motion representations for cross-id face reenactment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
12. Tang, H.; Wang, W.; Xu, D.; Yan, Y.; Sebe, N. Gesturegan for hand gesture-to-gesture translation in the wild. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Republic of Korea, 22–26 October 2018; pp. 774–782. [CrossRef]
13. Balakrishnan, G.; Zhao, A.; Dalca, A.V.; Durand, F.; Guttag, J. Synthesizing images of humans in unseen poses. In Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8340–8348. [CrossRef]
14. Wiles, O.; Koepke, A.S.; 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, Munich, German, 8–14 September 2018; pp. 670–686. [CrossRef]
15. Siarohin, A.; Lathuiliere, 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, Long Beach, CA, USA, 15–20 June 2019; pp. 2377–2386. [CrossRef]
16. Siarohin, A.; Lathuiliere, S.; Tulyakov, S.; Ricci, E.; Sebe, N. First order motion model for image animation. *Neural Inf. Process. Syst.* **2019**, *32*. [CrossRef]
17. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision: ECCV, Munich, German, 8–14 September 2018; pp. 3–19. [CrossRef]
18. Hornik, K.; Stinchcombe, M.; White, H. Multilayer feedforward networks are universal approximators. *Neural Netw.* **1989**, *2*, 359–366. [CrossRef]
19. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 July 2016; pp. 770–778. [CrossRef]
20. Nagrani, A.; Chung, J.S.; Zisserman, A. Voxceleb: A large-scale speaker identification dataset. *arXiv* **2017**, arXiv:1706.08612.
21. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7291–7299. [CrossRef]
22. Bulat, A.; Tzimiropoulos, G. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230, 000 3d facial landmarks). In Proceedings of the IEEE International Conference on Computer Vision: ICCV, Venice, Italy, 22–29 October 2017; pp. 1021–1030. [CrossRef]
23. Hermans, A.; Beyer, L.; Leibe, B. In defense of the triplet loss for person re-identification. *arXiv* **2017**, arXiv:1703.07737.
24. Amos, B.; Ludwiczuk, B.; Satyanarayanan, M. Openface: A general-purpose face recognition library with mobile applications. *CMU Sch. Comput. Sci.* **2016**, *6*, 20.
25. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]