IEICE TRANS. INF. & SYST., VOL.E106–D, NO.1 JANUARY 2023

PAPER Special Section on Enriched Multimedia–Advanced Safety, Security and Convenience–

Face Image Generation of Anime Characters Using an Advanced First Order Motion Model with Facial Landmarks

Junki OSHIBA[†a)], ***Nonmember***, **Motoi IWATA**[††b)], ***Senior Member***, ***and*** **Koichi KISE**[††c)], ***Fellow***

SUMMARY Recently, deep learning for image generation with a guide for the generation has been progressing. Many methods have been proposed to generate the animation of facial expression change from a single face image by transferring some facial expression information to the face image. In particular, the method of using facial landmarks as facial expression information can generate a variety of facial expressions. However, most methods do not focus on anime characters but humans. Moreover, we attempted to apply several existing methods to anime characters by training the methods on an anime character face dataset; however, they generated images with noise, even in regions where there was no change. The first order motion model (FOMM) is an image generation method that takes two images as input and transfers one facial expression or pose to the other. By explicitly calculating the difference between the two images based on optical flow, FOMM can generate images with low noise in the unchanged regions. In the following, we focus on the aspect of the face image generation in FOMM. When we think about the employment of facial landmarks as targets, the performance of FOMM is not enough because FOMM cannot use a facial landmark as a facial expression target because the appearances of a face image and a facial landmark are quite different. Therefore, we propose an advanced FOMM method to use facial landmarks as a facial expression target. In the proposed method, we change the input data and data flow to use facial landmarks. Additionally, to generate face images with expressions that follow the target landmarks more closely, we introduce the landmark estimation loss, which is computed by comparing the landmark detected from the generated image with the target landmark. Our experiments on an anime character face image dataset demonstrated that our method is effective for landmark-guided face image generation for anime characters. Furthermore, our method outperformed other methods quantitatively and generated face images with less noise.

key words: *comic computing, image generation, facial landmark, generative adversarial network*

1. Introduction

In recent years, with the spread of electronic devices, such as smartphones and tablets, we have had increasing opportunities to read comics. Additionally, comics with animation (motion comics) have appeared that take advantage of electronic devices that can play music and videos. We can read comics with animation, such as parallel shifts of anime characters, scaling of speech balloons, and swinging onomatopoeia. However, a limited number of comics can be enjoyed as motion comics because of the time and effort required to create them. The ultimate goal of this study is to generate motion comics from existing comics so that many comics can be enjoyed as motion comics.

In motion comics, the animation of characters' facial expressions appears frequently. Furthermore, it is more difficult to create facial expression change animation than other types of animation, such as parallel shifts of characters, scaling of speech balloons, and swinging onomatopoeia. Therefore, we focus on the generation of the animation of a character's facial expression.

Research on deep learning for image generation with a guide for the generation is progressing [1]–[9]. In most cases, researchers use a "source object" and "target pose," where the source object is an object to be animated and the target pose is a pose to be assigned to it. By applying image generation methods to a source object and the sequence of target poses, the corresponding animation is generated. The source object is usually an image and the target pose can be, for example, text, a vector, or an image.

A facial landmark is a more accurate way to represent target facial expressions. A facial landmark is a point on the contours of facial parts, such as the eyes and nose. Using facial landmarks, the shape and position of the eyes and mouth can be specified in detail, and a variety of facial expressions can be represented [3], [5], [6], [8], [10]–[14]. Additionally, facial landmarks can be automatically estimated from a face image by existing methods [10], [12]. In particular, Marco et al. defined the format of facial landmarks suitable for manga characters and proposed a method to estimate facial landmarks based on it [13]. Face image generation methods that use facial landmarks as the target pose have been proposed for human faces [3], [5], [6], [8], [14].

Siarohin et al. proposed the first order motion model (FOMM), which is an image generation method that can reduce noise in unchanged regions [4]. FOMM takes two images in the same category (e.g., faces, human bodies, etc.) but of different individuals as the source object and target pose. For example, if the source object is the face image of a person, then the target pose is the face image of a person different from the source object. If we are to apply FOMM to the generation of facial images of anime characters, the source object and target pose are supposed to be the face images of anime characters different from the sources, where we have already confirmed that the employment of human face images as targets causes low-quality generated images. We could find a target face image with an appropriate fa-

Manuscript received March 23, 2022.
Manuscript revised August 14, 2022.
Manuscript publicized October 12, 2022.

[†]The author is with Osaka Prefecture University, Sakai-shi, 599–8531 Japan.
[††]The authors are with Osaka Metropolitan University, Sakai-shi, 599–8531 Japan.
a) E-mail: oshiba@m.cs.osakafu-u.ac.jp
b) E-mail: imotoi@omu.ac.jp (Corresponding author)
c) E-mail: kise@omu.ac.jp

DOI: 10.1587/transinf.2022MUP0004

cial expression if we had a large dataset of anime characters with an appropriate copyright license agreement. Currently, there is no such dataset. Therefore, as an alternative plan, we introduce facial landmarks as the target pose. The establishment of this method will give general users without special skills to draw various facial expressions of anime characters the way to add a facial expression motion to still anime character images easily.

In this paper, we propose an advanced FOMM method to use an image in a different category from the source object as the target pose, where different category means anime character face images and facial landmark images. Our experiments demonstrated that the proposed method is effective for the face image generation of anime characters based on a character face landmark, where the source object is the face image of a character and the target pose is the character face landmark.

2. Related Work

To our best knowledge, all the state-of-the-art methods for face image generation are based on DNN. In this section, we introduce previous DNN methods related to ours. Since 2018, many face image generation methods have used a source object and target pose, where the source object is a face image [3]–[9], [14], [16]–[18]. Choi et al. proposed StarGAN, which uses text labels for facial expressions (happy, angry, etc.) as the target pose and generates face images with expressions based on the labels [16]. Hao et al. proposed C^{2}GAN, which uses facial landmarks as the target pose and generates face images with expressions based on the facial landmarks [14]. Their experimental results demonstrated that C^{2}GAN is effective for human face image generation with facial landmarks as a target pose. However, when C^{2}GAN is trained on the face images of anime characters, it generates images with noise, even in regions that do not change from the source object image [19]. Marco et al. defined the format of facial landmarks suitable for manga characters and proposed a method to estimate facial landmarks as a set of 60 points from a face image of a manga character [13]. Figure 2 shows the facial landmarks estimated by the method of Marco et al. plotted on a face image. Hereafter, the set of 60 facial landmarks defined by Marco et al. is called the "character face landmark."

Siarohin et al. proposed the FOMM, which is an image generation method that can reduce noise in unchanged regions by explicitly calculating the difference between the source object and target pose [4]. In the context of face image generation, FOMM takes two face images but of different individuals as the source object and target pose. Figure 1 shows an overview of FOMM. First, the motion module calculates difference information based on optical flow between the source and target images. Next, the generation module generates a face image with the expressions on the target image from the difference information and source image. The employment of the difference information reduces the noise in unchanged regions. In preliminary experiment,

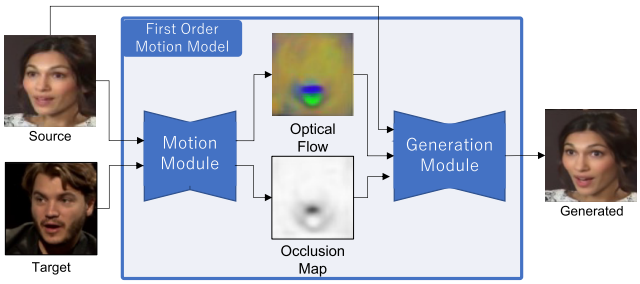


Fig. 1 Overview of the first order motion model [4].

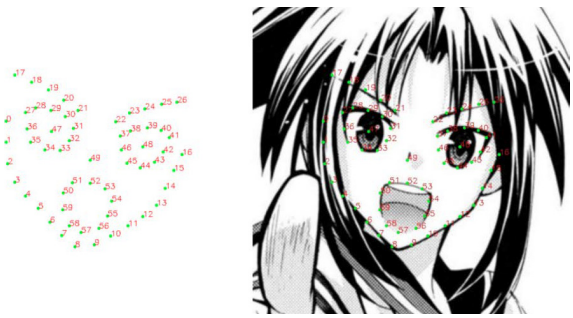


Fig. 2 Face image of a manga character with a landmark plotted in green dots.†

we checked the performance of FOMM when source image and target image are anime face image and human face image, respectively. The eyes of a target image were often matched to the upper parts of the eyes of the source image. It causes that only the upper parts of the eyes were changed in the generated image. Moreover, the nose of a target image was sometimes matched to the mouth of the source image. It may be because the nose of the source image did not have enough characteristics to be matched to the nose of the target image.

Currently, the above methods have been shown to be ineffective in generating face images of anime characters. Zhang et al. proposed CPTNet, which is a method for generating face images of anime characters [9]. CPTNet takes a single character's face image as a source object and a pose vector that represents the face pose and head pose as the target pose. CPTNet uses a pose vector that represents the face pose and head pose as the target pose and generates a face image with the face and head pose based on the pose vector. One disadvantage of CPTNet is that the animation of facial expression changes generated by CPTNet is limited to blinking and simple vertical opening and closing of the mouth. It is because the only parameters related to facial expressions in the pose vector are the eyes and mouth openings.

Our contribution in this paper is the improvement of FOMM so that the images in different categories can be used as source and target images, where the images in different categories are anime character face images and facial landmark images. The core idea of the contribution is the introduction of a landmark estimator and landmark estimation loss to FOMM.

†©Takuji in Manga109 dataset [15]

3. Proposed Method

In this section, we describe our proposed method for generating a facial image with an expression following a character face landmark.

3.1 Overview

Figure 3 shows an overview of the training flow of our proposed method. To avoid copyright issues, we use the illustrations of Tohoku Zunko† in Fig. 3, Fig. 4, Fig. 5, Fig. 6, and Fig. 7, where they can be used for non-commercial purposes without permission. The inputs are the face image of an anime character as the source object (called "source image," denoted by S), a character face landmark estimated from the source image (called "source landmark," denoted by L_S), and a character face landmark used as the target pose (called "target landmark," denoted by L_T). In Fig. 3, T, $\hat{T}$, and $\hat{L_T}$ represent ground truth images, the corresponding generated images, and the estimated landmark images from the generated images, respectively. As can be seen by comparing Fig. 1 and Fig. 3, the motion module and the generation module are the same as the original FOMM. First, the motion module generates the difference information based on optical flow between the source and target landmarks. The generation module then takes the difference information and source image S as inputs and generates a face image $\hat{T}$ which is expected to present the target pose L_T. During training, a character face landmark is estimated from the generated images, and the landmark estimation loss is calculated in addition to the existing losses of FOMM.

We describe the introduction of the character face landmark to FOMM in Sect. 3.2 and the landmark estimation loss in Sect. 3.3.

3.2 Introduction of Facial Landmarks to the First Order Motion Model

To use a character face landmark, we convert it into an image. We call the imaged character face landmark the "landmark image." As shown in Fig. 2, each point in a landmark has a label corresponding to its position. For example, the 17th point represents the right edge of the right eyebrow of the character. In our method, we generate the landmark image by filling in each part of the face with different colors based on the labels. Figure 4 shows an example to obtain a landmark image. First, the facial landmark (Fig. 4 (b)) is estimated from the face image (Fig. 4 (a)). After that, the landmark image (Fig. 4 (c)) is obtained by connecting the points of each part and filling them in Fig. 4 (b).

The simplest approach to introducing the landmark image to FOMM is to use the landmark image as the target pose instead of a face image. In this approach, the source image and target landmark image are input into the motion module. However, inappropriate difference information based on optical flow would be generated from a face image and landmark image. Therefore, we change the inputs to FOMM to the source image, source landmark image, and target landmark image. The motion module generates the difference information based on optical flow from the source and target landmark images. The generation module takes the difference information and source image as inputs, and is supposed to generate a facial image from which we can obtain the target landmark image. Moreover, we add a landmark estimator to FOMM to calculate the landmark estimation loss. We call FOMM with the above improvements "advanced FOMM."

†©2022 SSS LLC., https://zunko.jp/

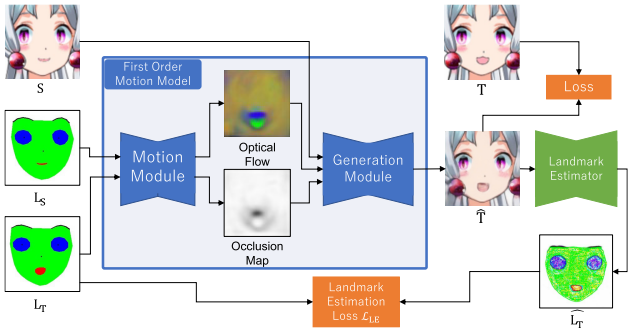


Fig. 3 Training flow of the proposed method. The inputs for training are source image S, source landmark image L_S, target landmark image L_T, and target image T. T, $\hat{T}$, and $\hat{L_T}$ represent ground truth images, the corresponding generated images, and the estimated landmark images from the generated images, respectively.

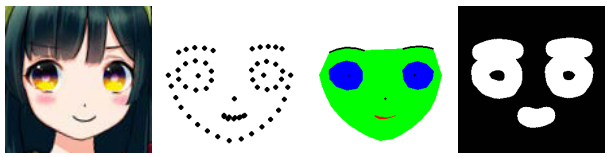


Fig. 4 Example of the landmark, landmark image, and landmark masks. (a) Face image, (b) facial landmark estimated by the method of Marco et al., (c) landmark image, and (d) landmark mask.

3.3 Landmark Estimation Loss

We define the landmark estimation loss, which is the error between a character face landmark from an image generated by the generation module and the target landmark, inspired by C^{2}GAN. The purpose is to generate face images with expressions that follow target landmarks more closely.

3.3.1 Landmark Estimator

The landmark estimator generates a landmark image from an image generated by the generation module. The implementation of the landmark estimator is based on the keypoint generator of C^{2}GAN, and uses U-net [20].

3.3.2 Calculation of Landmark Estimation Loss

We calculate the landmark estimation loss by comparing the estimated landmark image with the target landmark image.

We consider the comparison region of the landmark images. If the landmark estimation loss is calculated by comparing the entire landmark images, the landmark estimation loss is affected by errors in regions that are not related to facial expression changes, such as those outside the face. To address this problem, we use a landmark mask to calculate the landmark estimation loss. The landmark mask is a binary representation of the regions around the contours of the eyebrows, eyes, mouth, and other regions, as shown in Fig. 4 (d), and we generate it by connecting the facial landmark dots that represent the eyebrows, eyes, and mouth among the facial landmarks with thick lines. We calculate the landmark estimation loss $\mathcal{L}_{\mathrm{LE}}$ using the following equation:

$$\mathcal{L}_{\mathrm{LE}}(\hat{\mathrm{L}}_{\mathrm{T}}, \mathrm{L}_{\mathrm{T}}) = \frac{1}{N} \sum_{(h,w)\in \boldsymbol{M}} \sum_{c\in C} \left\{ \hat{\mathrm{L}}_{\mathrm{T}}(h,w,c) - \mathrm{L}_{\mathrm{T}}(h,w,c) \right\}^2, \quad (1)$$

where $\boldsymbol{M}$ denotes the landmark mask, N denotes the area of the landmark mask (the area of the white region in Fig. 4 (d)), $C = \{R, G, B\}$ denotes the set of color components of the image, and $\mathrm{L}_{\mathrm{T}}(h, w, c)$ denotes the value of the h-th row, w-th column, and color component c in the landmark image L_{T}.

We calculate the total loss of our method by adding weighted $\mathcal{L}_{\mathrm{LE}}(\hat{\mathrm{L}}_{\mathrm{T}}, \mathrm{L}_{\mathrm{T}})$ to the existing loss of FOMM $\mathcal{L}_{\mathrm{FOMM}}(\hat{\mathrm{T}}, \mathrm{T})$, which we calculate from the generated image $\hat{\mathrm{T}}$ and ground truth image T. We calculate the total loss $\mathcal{L}_{\mathrm{all}}(\hat{\mathrm{T}}, \mathrm{T}, \hat{\mathrm{L}}_{\mathrm{T}}, \mathrm{L}_{\mathrm{T}})$ using the following equation:

$$\mathcal{L}_{\mathrm{all}}(\hat{\mathrm{T}}, \mathrm{T}, \hat{\mathrm{L}}_{\mathrm{T}}, \mathrm{L}_{\mathrm{T}}) = \mathcal{L}_{\mathrm{FOMM}}(\hat{\mathrm{T}}, \mathrm{T}) + \lambda \mathcal{L}_{\mathrm{LE}}(\hat{\mathrm{L}}_{\mathrm{T}}, \mathrm{L}_{\mathrm{T}}), \quad (2)$$

where λ is weight of $\mathcal{L}_{\mathrm{LE}}(\hat{\mathrm{L}}_{\mathrm{T}}, \mathrm{L}_{\mathrm{T}})$.

3.4 Dataset Preparation

In this study, we prepared the dataset using anime videos to obtain a large amount of data. Before describing the generation procedure, we describe the data required for training the proposed method. Our proposed method requires two face images and their character face landmarks that satisfy the condition that the expressions of the same anime character are different (condition A). Additionally, the proposed method is not designed to generate outside the face (e.g., background, clothing, hairstyle) because a character face landmark only has information about the face region. Therefore, it is desirable for the two face images with different expressions to satisfy the condition that there are no changes outside the face (condition B), in addition to condition A.

First, we group images that satisfy condition A and, to some extent, condition B using shot division, as described in Sect. 3.4.1. Second, we detect and crop faces as described in Sect. 3.4.2. Third, we estimate the character face landmark from each face image, as described in Sect. 3.4.3. Fourth, we normalize the face images and character face landmarks, as described in Sect. 3.4.4. Finally, we pair images within each group divided using shot division, as described in Sect. 3.4.5.

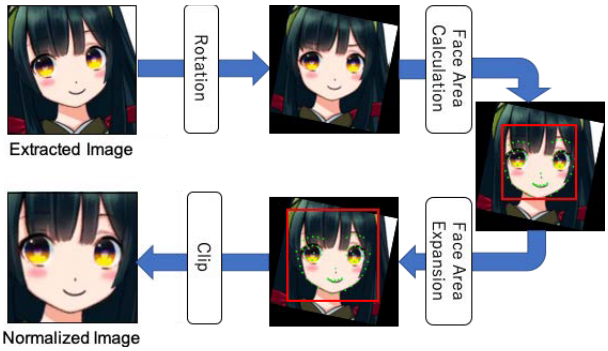


Fig. 5 Normalization procedure.

3.4.1 Shot Division

A shot is one of the units of a video and is a sequence of motion pictures taken by a single camera. The background, clothing, and hairstyle tend not to change significantly in the same shot. Therefore, two images extracted from the same shot are stored in the same group, assuming that they satisfy conditions A and B. The shot division is automatically performed based on the similarity of color histograms to reduce the labor. After the next step described in Sect. 3.4.2, the errors of the shot division will be corrected manually.

3.4.2 Face Detection

The face regions are detected and cropped from the shots to obtain face images using OpenCV[†]. Then, the shots without face regions are deleted. The obtained face images are stored in a group for each shot.

3.4.3 Facial Landmark Estimation

The character face landmark is detected from the face images using the method of Marco et al. [13]. Although the method of Marco et al. is applied to monochrome comic images, the method can also detect character face landmarks for color anime images.

3.4.4 Normalization

Figure 5 shows the normalization procedure for aligning the position, size, and orientation of the character face in all face images. First, the face image is rotated so that the lines passing through the center of gravity of each eye are horizontal. In this case, the area outside the image is padded by black pixels. Next, the square face area is calculated based on the left-most, right-most, and bottom-most points in the contour landmark. The normalized image is cropped by extending the 10% face area of one side of the square. Additionally, the character face landmarks are normalized to correspond to the normalized images.

[†]https://github.com/nagadomi/lbpcascade_animeface

3.4.5 Generating Pairs

Before generating pairs, all images are visually checked. Images that are assessed to have failed in face detection or facial landmark detection are removed. Because any combination of images in the same group satisfies conditions A and B, all combinations other than exact same images are selected as pairs.

4. Experiment

4.1 Dataset

In this experiment, we used a dataset prepared from anime videos using the procedure described in Sect. 3.4. The dataset consisted of 6,897 pairs of facial expression changes of 284 anime characters from 89 anime titles, where the 89 anime titles were in a variety of drawing styles and production teams. We divided the 6,897 pairs into five sets and used five-fold cross-validation, in which one set was used for testing and the remaining four sets were used for training. When dividing the dataset, we assigned one of the five sets to each character so that pairs of the same character would belong to the same set.

4.2 Evaluation Protocol

To quantitatively evaluate the quality of the generated images, we used the peak signal-to-noise ratio (PSNR), structural similarity (SSIM) [21], and learned perceptual image patch similarity (LPIPS) [22], which are described below.

PSNR is calculated based on the pixel value error at the same coordinate. The larger the value, the closer the generated image to the ground truth.

SSIM is calculated using a structural_similarity function. The larger the value, the closer the generated image to the ground truth. Image quality evaluation by SSIM is closer to human perception than that by PSNR. For example, if an entire image is shifted, the PSNR value will decrease significantly, while the SSIM value will decrease gently.

LPIPS is an evaluation metric calculated by inputting the generated image and ground truth image into the same trained convolutional neural network and comparing the features obtained by convolution. Experiments by Zhang et al. demonstrated that LPIPS is closer to human perception than PSNR or SSIM [22]. For example, when we compare an image with mild noise with the image with strong blur, both PSNR and SSIM values of the image with strong blur are higher, while LPIPS value of the image with mild noise is higher. The smaller the value, the closer the generated image to the ground truth image.

The character face landmarks used in the proposed method do not contain information outside the face region. Thus, it is difficult to generate the background and hair regions appropriately, and the performance of face region generation cannot be correctly evaluated. Therefore, we also calculated PSNR, SSIM, and LPIPS for the face region. The face region is defined by connecting the facial contour and the landmarks of both eyebrows. We calculated PSNR only from the values of pixels in the face region, regardless of the shape of the face region. By contrast, because SSIM and LPIPS are complicated to implement, we filled the area outside the face region with white and calculated SSIM and LPIPS for the rectangular region circumscribed to the face region. We computed the mouth and eye regions of the rectangle using landmarks, and also calculated PSNR, SSIM, and LPIPS for each region.

4.3 Ablation Study

Ablation study aims at investigating the performance of AI systems by removing certain components to understand the contribution of the component to the overall system. To show the effectiveness of changing the input into FOMM and adding the landmark estimation loss, we compare models trained under different conditions as follows:

Base: a model that inputs a source image and target landmark image to the motion module in FOMM.

LL+$\mathcal{L}_{\mathrm{LE}}$ model: a model that inputs source and target landmark images to the motion module in FOMM and adds a landmark estimator to calculate the landmark estimation loss. We experimentally set λ in Eq. (2) to 1.

Landmark-Landmark model (LL model): a model where the landmark estimator is removed from LL+$\mathcal{L}_{\mathrm{LE}}$ model.

Image-Image model (II model): a model applying FOMM to the face image generation of anime characters. For this model, we did not use character face landmarks. We input the source and target face images into the motion module.

The II model uses a face image as the target pose, whereas the Base model, LL model, and LL+$\mathcal{L}_{\mathrm{LE}}$ model use a landmark image as the target pose. We can say that the II model has an advantage over the others, in the sense that only the II model takes the target image to be generated, that is, the II model is told the answers while the others are not. Although these differences exist in the experimental settings, we added the II model to the comparison to verify the difference in performance when a landmark image is used as the target and when a face image is used as the target.

For all the four models, we have set the number of epochs to 200 and batch size to 36. These values are the same as those for FOMM.

Table 1 shows the quantitative evaluation results. Comparing the LL model with the Base model, the LL model was superior for all 12 evaluation metrics. Focusing on the face region, the LL model improved Face-PSNR by 1.0562 dB, Face-SSIM by 0.0354, and Face-LPIPS by 0.0324. There-

Table 1 Quantitative evaluation of the ablation study.

	All			Face			Mouth			Eye		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Base model	16.4073	0.5350	0.2657	20.0644	0.6999	0.1984	22.4518	0.7351	0.1889	15.2242	0.4521	0.1764
LL model	17.2773	0.5743	0.2285	21.1206	0.7353	0.1660	23.9281	0.7795	0.1444	16.2933	0.5148	0.1550
LL+$\mathcal{L}_{\mathrm{LE}}$ model	**17.3071**	**0.5755**	**0.2282**	**21.1648**	**0.7364**	**0.1652**	**24.0117**	**0.7824**	**0.1409**	**16.3445**	**0.5174**	**0.1540**
II model	18.3375	0.6102	0.2168	22.1510	0.7572	0.1578	23.9584	0.7681	0.1572	17.2448	0.5517	0.1513

Fig. 6 Qualitative evaluation of the ablation study.

fore, the results demonstrated that it was effective to change the input to the motion module to source and target landmark images instead of simply replacing the target pose from a face image with a landmark image.

Comparing the LL model and LL+$\mathcal{L}_{\mathrm{LE}}$ model, all 12 evaluations improved, although the numerical improvements were smaller than those that resulted from a comparison of the Base model and LL model. Although Face-LPIPS improved by only 0.0008, Mouth-LPIPS improved by 0.0035, which indicates that the addition of the landmark estimation loss was particularly effective in generating mouth regions.

The comparison of the LL model and II model is described as follows: Note that the II model has an advantage over the LL model because it uses the ground truth face image as the target pose during testing. In terms of LPIPS, where smaller values indicate better performance, the LL model had 0.0117 larger All-LPIPS, 0.0082 larger Face-LPIPS, and 0.0037 larger Eye-LPIPS than the II model. By contrast, the LL model had 0.0128 smaller Mouth-LPIPS than the II model, which indicates that using landmark images as the target pose was effective in generating mouth regions.

Figure 6 shows examples of the generated images for each model. The row (a) in Fig. 6 is the examples of a mouth opening. We can see that the LL and LL+$\mathcal{L}_{\mathrm{LE}}$ models could successfully generate images where the mouth is widely opened. The row (b) in Fig. 6 is the examples of eyes closing. For this case, both the LL and LL+$\mathcal{L}_{\mathrm{LE}}$ models succeeded to generate images where eyes are closed. The row (c) in Fig. 6 is the examples of eyes and mouth closing. Only the LL+$\mathcal{L}_{\mathrm{LE}}$ model had successfully generated an image where both eyes and mouth are closed. Focusing on the opening and closing of the mouth, as shown in row (a), images that changed not only vertically but also horizontally were generated. Furthermore, as shown in rows (b) and (c),

Table 2 Quantitative comparison with existing methods, where "Ours" are the results on LL+$\mathcal{L}_{\mathrm{LE}}$ model.

	All			Face			Mouth			Eye		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
C^2GAN	15.2851	0.4617	0.3633	18.9814	0.6623	0.2617	22.2131	0.7270	0.2351	13.8600	0.3726	0.2481
bi-layer model	14.5997	0.4340	0.4196	18.3461	0.6462	0.3052	21.8795	0.7210	0.2647	13.2571	0.3392	0.2907
Ours	**17.3071**	**0.5755**	**0.2282**	**21.1648**	**0.7364**	**0.1652**	**24.0117**	**0.7824**	**0.1409**	**16.3445**	**0.5174**	**0.1540**

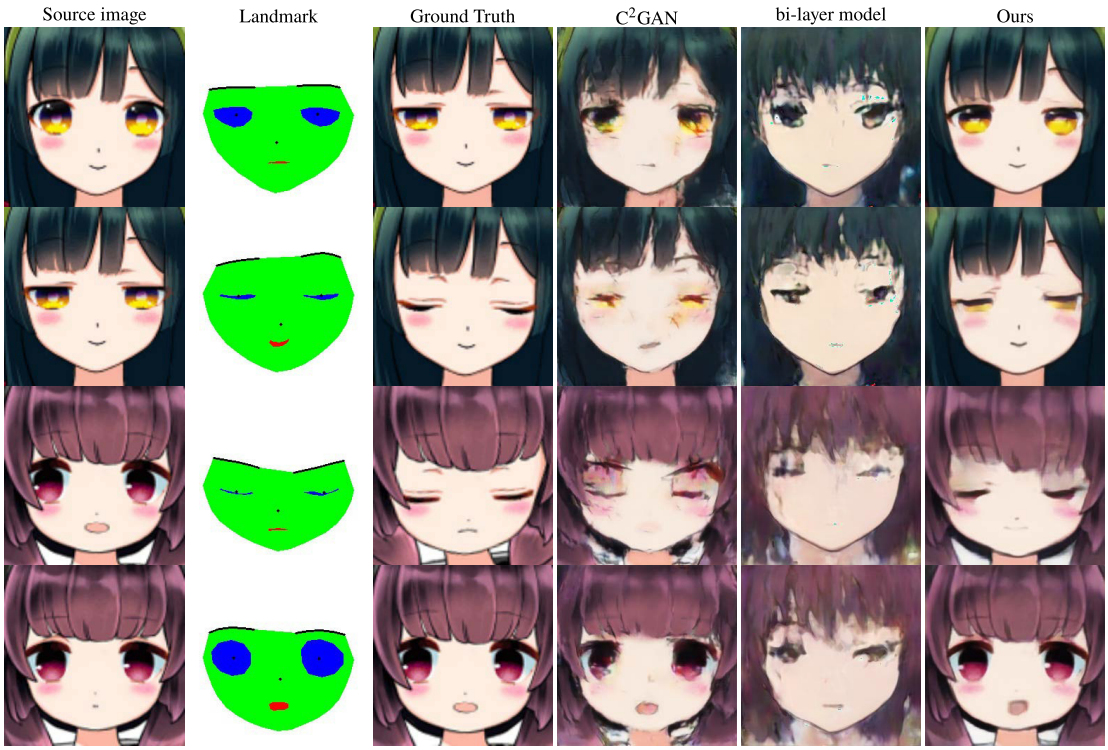


Fig. 7 Qualitative comparison with existing methods, where "Ours" are the results on LL+$\mathcal{L}_{\mathrm{LE}}$ model.

the LL+$\mathcal{L}_{\mathrm{LE}}$ model generated images with fully closed eyes in the case in which the Base model failed.

As shown in row (a), it was possible to make the mouth open, but in the case of the eyes opening, as shown in row (d), an image with a different eye style from the ground truth image was generated. It was difficult to generate open eyes from closed eyes for anime characters because of the large differences in eye styles among the anime characters.

Row (e) is an example in which the corners of the mouth were changed from up to down. However, the generated image was strongly influenced by the source image, and the corners of the mouth were up. Although the proposed method could generate rough shape changes, such as opening and closing the mouth, in the case of detailed shape changes, such as the raising and lowering of the corners of the mouth, it tended to generate images whose shapes were close to the source image, regardless of the target landmark image.

In the comparison of the LL+$\mathcal{L}_{\mathrm{LE}}$ model and II model, the former was inferior in the quantitative evaluation, but there was no obvious qualitative degradation.

4.4 Comparison with Existing Methods

We trained two existing face image generation methods, C^2GAN and the bi-layer model proposed by Zakharov et al. [8], which use facial landmarks as the target pose, on the same dataset and compared them with our method. Both C^2GAN and the bi-layer model use facial landmarks as images. Each method uses a different approach to draw facial landmarks. Therefore, in the preliminary experiment, we checked the most suitable approach to drawing facial landmarks for C^2GAN and the bi-layer model. Based on the Face-LPIPS value, we used the same approach as our method for C^2GAN and the original approach for the bi-layer model.

Table 2 shows the quantitative evaluation results for each method. Our method in Table 2 is the LL+$\mathcal{L}_{\mathrm{LE}}$ model in Table 1. Our method outperformed the existing methods for all evaluation metrics.

Figure 7 shows examples of the generated images for each existing method. The results demonstrated that our method generated images with low noise and clear contours, whereas the other methods produced severely distorted im-

ages. The effectiveness of our method in landmark-based face image generation for anime characters was demonstrated.

5. Conclusion

In this paper, we have presented an advanced FOMM method where the inputs to the motion module has been changed and landmark estimation loss is added so that the pose of a source anime character face image could be transformed using a target facial landmark image although their difference information based on optical flow cannot be calculated directly. In the experiment using the dataset generated from anime videos, our method improved the PSNR by 0.8998 dB, SSIM by 0.0405, and LPIPS by 0.0375 compared with a simple method that used facial landmarks using FOMM. Furthermore, compared with existing landmark-based face image generation methods, our method generated images with less noise and clearer contours. In the experimental results, even LL+$\mathcal{L}_{\mathrm{LE}}$ model could not generate opened eyes from completely closed eyes appropriately. Moreover, some slight changes on mouth were not reflected to the generated images. For practical use, such problems should be resolved. In addition to them, the proposed method does not work for characters with different facial landmarks from "character face landmark," for example, non-human characters (with three eyes, beast-men, etc.). The flexibility of facial landmarks is another issue to be solved for practical use. In this paper, we experimented only with face images of anime characters. However, our method is applicable to full-body pose transformation using full-body images and skeleton information.

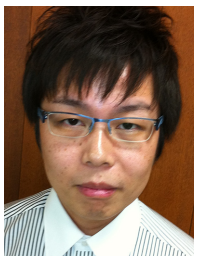
Acknowledgments

This work was supported in part by grants from JSPS Grant-in-Aid for Scientific Research (B) (Grant No. 20H04213), and JST CREST (Grant No. JPMJCR16E1). We thank Maxine Garcia, PhD, from Edanz for editing a draft of this manuscript (Edanz: https://jp.edanz.com/ac).

References

[1] A. Pumarola, A. Agudo, A.M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "Ganimation: Anatomically-aware facial animation from a single image," Proc. European Conference on Computer Vision (ECCV), pp.818–833, 2018.
[2] C. Chan, S. Ginosar, T. Zhou, and A.A. Efros, "Everybody dance now," Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019.
[3] W. Wang, X. Alameda-Pineda, D. Xu, P. Fua, E. Ricci, and N. Sebe, "Every smile is unique: Landmark-guided diverse smile generation," Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp.7083–7092, 2018.
[4] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," Advances in Neural Information Processing Systems, vol.32, pp.7137–7147, 2019.
[5] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, "Few-shot adversarial learning of realistic neural talking head models," Proc. IEEE/CVF Int. Conf. Comput. Vis., pp.9459–9468, 2019.
[6] T.C. Wang, A. Mallya, and M.Y. Liu, "One-shot free-view neural talking-head synthesis for video conferencing," Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021.
[7] O. Wiles, A.S. Koepke, and A. Zisserman, "X2face: A network for controlling face generation using images, audio, and pose codes," Proc. European Conference on Computer Vision (ECCV), pp.670–686, 2018.
[8] E. Zakharov, A. Ivakhnenko, A. Shysheya, and V. Lempitsky, "Fast bi-layer neural synthesis of one-shot realistic head avatars," European Conference on Computer Vision (ECCV), Aug. 2020.
[9] J. Zhang, K. Xian, C. Liu, Y. Chen, Z. Cao, and W. Zhong, "CPTNet: Cascade pose transform network for single image talking head animation," Proc. Asian Conference on Computer Vision, 2020.
[10] M. Kowalski, J. Naruniec, and T. Trzcinski, "Deep alignment network: A convolutional neural network for robust face alignment," Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, pp.88–97, 2017.
[11] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," CMU School of Computer Science, vol.6, 20 pages, 2016.
[12] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2d pose estimation using part affinity fields," IEEE Trans. Pattern Anal. Mach. Intell., vol.43, no.1, pp.172–186, Jan. 2021.
[13] M. Stricker, O. Augereau, K. Kise, and M. Iwata, "Facial landmark detection for manga images," arXiv preprint arXiv:1811.03214, 2018.
[14] H. Tang, D. Xu, G. Liu, W. Wang, N. Sebe, and Y. Yan, "Cycle in cycle generative adversarial networks for keypoint-guided image generation," ACM MM, pp.2052–2060, Oct. 2019.
[15] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki, and K. Aizawa, "Sketch-based manga retrieval using manga109 dataset," Multimed. Tools. Appl., vol.76, no.20, pp.21811–21838, 2017.
[16] Y. Choi, M. Choi, M. Kim, J.W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp.8789–8797, 2018.
[17] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "Animating arbitrary objects via deep motion transfer," Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp.2377–2386, 2019.
[18] S. Tulyakov, M.Y. Liu, X. Yang, and J. Kautz, "Mocogan: Decomposing motion and content for video generation," Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp.1526–1535, 2018.
[19] J. Oshiba, M. Iwata, and K. Kise, "Automatic landmark-guided face image generation for anime characters using C^2GAN," Int. Conf. Pattern Recognit., pp.236–249, Springer, 2021.
[20] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," Int. Conf. Medical Image Computing and Computer-Assisted Intervention, pp.234–241, Springer, 2015.
[21] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Trans. Image Process., vol.13, no.4, pp.600–612, April 2004.
[22] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp.586–595, 2018.

Junki Oshiba received the B.S. and M.E. degrees in computer and systems sciences from Osaka Prefecture University, Osaka, Japan, in 2020 and 2022, respectively. His research interests include comic computing and image generation.



Motoi Iwata received the M.E. and D.E. degrees in computer and systems sciences from Osaka Prefecture University, Osaka, Japan, in 1999 and 2005, respectively. He is now an associate professor of the Department of Core Informatics, Graduate School of Informatics, Osaka Metropolitan University, Japan. His current research focuses on digital watermarking, data hiding, image retrieval, comic computing, and educational technology. He is a member of the Imaging Society of Japan, the Institute of Image Information and Television Engineers, ACM, and IEEE.

Koichi Kise received the B.E., M.E., and Ph.D. degrees in communication engineering from Osaka University, Osaka, Japan, in 1986, 1988, and 1991, respectively. From 2000 to 2001, he was a visiting professor at the German Research Center for Artificial Intelligence (DFKI), Germany. He is now a professor of the Department of Core Informatics, Graduate School of Informatics, Osaka Metropolitan University, Japan. He received awards including the best paper award of IEICE in 2008 and 2022, the IAPR/ICDAR best paper awards in 2007 and 2013, the IAPR Nakano award in 2010, the ICFHR best paper award in 2010 and the ACPR best paper award in 2011. He worked as the chair of the IAPR technical committee 11 (reading systems), a member of the IAPR conferences and meetings committee. He is an editor-in-chief of the international journal of document analysis and recognition. His major research activities are in analysis, recognition and retrieval of documents, images and human activities. He is a member of IEEE, ACM, IPSJ, IEEJ, ANLP and HIS.