

## Research Article

# Face Swapping: Realistic Image Synthesis Based on Facial Landmarks Alignment

Dongyue Chen <sup>1</sup>, Qiusheng Chen <sup>1</sup>, Jianjun Wu <sup>1</sup>, Xiaosheng Yu <sup>2</sup>, and Tong Jia <sup>1</sup>

<sup>1</sup>College of Information Science and Engineering, Northeastern University, China

<sup>2</sup>Faculty of Robot Science and Engineering, Northeastern University, China

Correspondence should be addressed to Tong Jia; [jiatong@ise.neu.edu.cn](mailto:jiatong@ise.neu.edu.cn)

Received 30 November 2018; Revised 3 February 2019; Accepted 25 February 2019; Published 14 March 2019

Academic Editor: Bogdan Smolka

Copyright © 2019 Dongyue Chen et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

We propose an image-based face swapping algorithm, which can be used to replace the face in the reference image with the same facial shape and features as the input face. First, a face alignment is made based on a group of detected facial landmarks, so that the aligned input face and the reference face are consistent in size and posture. Secondly, an image warping algorithm based on triangulation is presented to adjust the reference face and its background according to the aligned input faces. In order to achieve more accurate face swapping, a face parsing algorithm is introduced to realize the accurate detection of the face-ROIs, and then the face-ROI in the reference image is replaced with the input face-ROI. Finally, a Poisson image editing algorithm is adopted to realize the boundary processing and color correction between the replacement region and the original background, and then the final face swapping result is obtained. In the experiments, we compare our method with other face swapping algorithms and make a qualitative and quantitative analysis to evaluate the reality and the fidelity of the replaced face. The analysis results show that our method has some advantages in the overall performance of swapping effect.

## 1. Introduction

Face synthesis refers to the image processing technology of the automatic fusion of two or more different faces into one face, which is widely used in fields of video synthesis, privacy protection, picture enhancement, and entertainment applications. For example, when we want to share some of the interesting things on social networks, we can use the face synthesis technique which can be regarded as a fusion of facial features and details to change our appearances appropriately without privacy leaks. As another type of face fusion, face swapping combines some parts of one person's face with other parts of the other's face to form a new face image. For instance, in the application of virtual hairstyle visualization, the client's facial area can be fused with the hair areas of the model images to form new photos, so that customers can virtually browse their own figures with different hairstyles. This paper focuses on the face swapping problem of virtual browsing applications for hairstyle and dressing. Our main contributions of the proposed algorithm include the following: (1) construct a pipeline of face swapping which integrates some

learning-based modules into the traditional replacement-based approach, (2) improve the sense of reality and reliability of the synthesis face based on the precise detection of the facial landmarks, and (3) the face occlusion problem can be solved by introducing an accurate face parsing algorithm.

## 2. Related Work

Existing face swapping algorithms can be roughly divided into three categories: replacement-based, model-based, and learning-based. The replacement-based method usually replaces the face region in the reference image with the input face region and then applies some image processing techniques to enhance the real sense of the synthesized image. Bitouk et al. [1] built a face image set with different expressions, facial shapes, and postures, and the algorithm could automatically select the most similar reference face from the image set for face replacement. Although this method can reduce the influence of the differences in facial expression, face shape, and posture, the application of

the algorithm is limited because users cannot choose the favorite reference face image independently. Mahajan et al. [2] established a mask image of facial features, and the mask-covered image was then attached to the reference image to achieve the face swapping. But the selected mask image, which can only cover the facial features, is unable to retain other facial characteristics, such as wrinkles, textures, skin colors, and muscle deformations. Besides, this method requests users to carry with the interactive instructions when running, which greatly reduces the practicability of the algorithm.

In the model-based approach [3], a two-dimensional or three-dimensional parametric feature model is established to represent human face, and the parameters and features are well-adjusted to the input image. Then the face reconstruction is performed on the reference image based on the result of adjusting the model parameters. An early work presented by Blanz and Volker et al. [4] used a 3D model to estimate the face shape and posture, which improved the shortcoming of the unsatisfied performance of the synthesis due to the illumination and the perspective. However, the algorithm requires a 3D input model and a manual initialization to get a better result, which undoubtedly has a stricter requirement for data acquisition. Wang et al. [5] proposed an algorithm based on active apparent model (AAM). By using the well trained AAM, the face swapping is realized in two steps: model fitting and component composite. But this method needs to specify the face-ROI manually and a certain number of face images for model training. Lin et al. [6] presented a method of constructing a 3D model based on the frontal face image to deal with the different perspectives of reference image and input image. But the reconstructed model does not reflect the characteristics of the original face precisely and takes too much time to compute.

In most of the learning-based models, the reference image is converted into a synthesized face image by training a generative neural network that contains the information of the input image. Korshunova et al. [7] proposed a model based on convolution neural network, which can change the reference face into the input face while maintaining the posture, expression, and illumination of the former one. Although this method [7] has some advantages in the sense of reality, it needs a lot of training data and a large amount of computation, and the trained network only works for one single person. In addition, gan (generative adversarial networks) technology can also be used to obtain the generation of face synthetic images [8]. This method replaces the latent space representation of the face and then reconstructs the entire face image with a region-separated gan model. However, this model needs to build a large face image dataset with pixel-level parsing labels while slightly degrading the quality of the synthetic picture.

Above all, the replacement-based approach is simple and fast but sensitive to the variation in posture and perspective. The model-based method can effectively solve the perspective problem; however, it usually needs to collect three-dimensional face data, and robustness is not something to be satisfied. The learning-based approach can produce quite real and natural synthetic face image, while usually

requiring a large number of training data and having more restrictions on the input and reference faces. Based on the comprehensive consideration of the characteristics of the above three methods, a face swapping algorithm supported by the facial landmark alignment is proposed under the replacement-based framework.

In addition, other widely used algorithms have been applied in our methods to achieve better results, such as facial landmark detection [9, 10], facial region segmentation [11, 12], and face warping [13]. And we will detail how these algorithms are applied in the method section.

### 3. Method

The algorithm is composed of three steps: face alignment, warping and replacement. The accuracy and robustness of the algorithm are enhanced by introducing some learning-based modules like facial landmark detection and face parsing.

*3.1. Pipeline of Face Swapping.* Although face swapping seems uncomplicated and practicable, an elaborately designed algorithm flow still has an impact on the realization of the final result. The pipeline of the proposed algorithm starts with two channels that finally fuse into one, as shown in Figure 1. First, the input image is aligned with the reference image based on a facial landmark detection algorithm. Second, the reference image is warped to fit the aligned face of the input image. With an advanced face parsing algorithm, in the next step, the face-ROIs are extracted from the aligned input image and the warped reference image, respectively. Finally, some common steps of face replacement and color correction are introduced to generate the final composite face image. To summarize, the proposed algorithm will be demonstrated into three parts: face alignment, face warping, and face replacement.

*3.2. Face Alignment.* As the first step of face swapping, alignment refers to aligning the input face image  $I_{in}$  and the reference face image  $I_{ref}$  in size and direction. For the purpose of detecting faces in pictures, we apply the relevant methods in paper [14] which proposes a novel multiple sparse representation framework for visual tracking to detect the faces in pictures. Apart from increasing the speed of the algorithm, the application of sparse coding and dictionary learning also enables these methods to learn more knowledge from relatively fewer sample data. Then we extract several stable key points from the images to mark the faces, referred to as facial landmark detection (FLD for short). In this paper, we employ a popular FLD algorithm [9, 15] based on an ensemble of regression trees to detect facial landmarks  $\Omega_s = \{s_i \mid i = 1, 2, \dots, 68\}$ , as plotted in Figure 2(a). We use this method that relies on our implementation. Each landmark point  $s_i$  is symmetric to another point  $s_{i'}$  with respect to the central axis of the face such as  $s_{22}$  and  $s_{23}$ ,  $s_{49}$ , and  $s_{55}$ . The points located at the central axis are symmetric to themselves such as  $s_{28}$  and  $s_{31}$ . To evaluate the rotation between the input and the reference face, the central axis of the input face should be extracted previously. According to the basic definition,

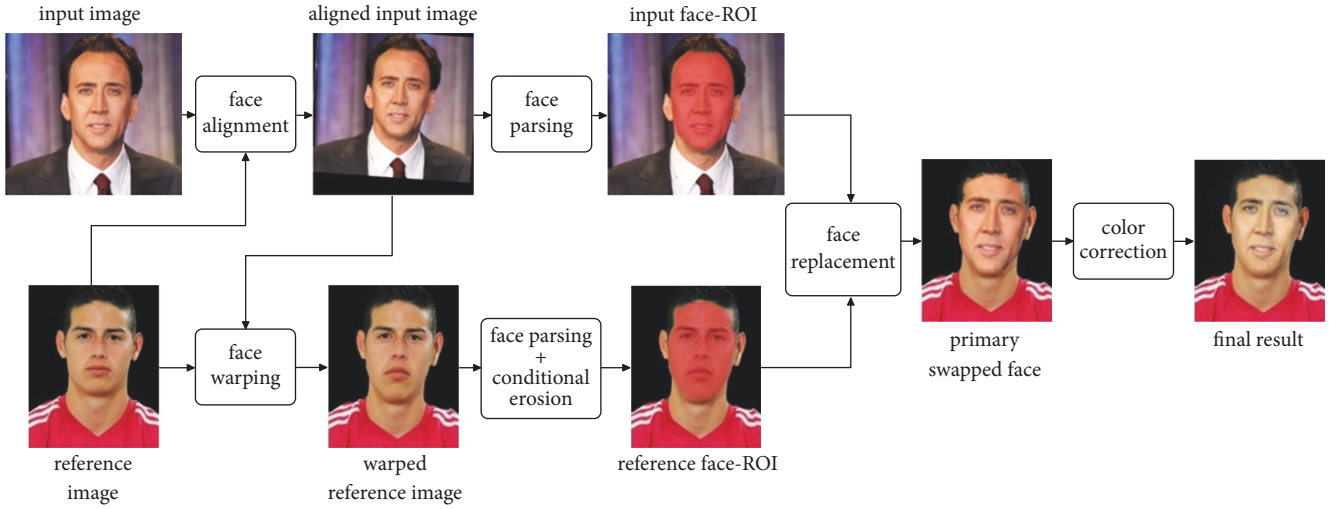


FIGURE 1: Pipeline of the proposed face swapping algorithm.

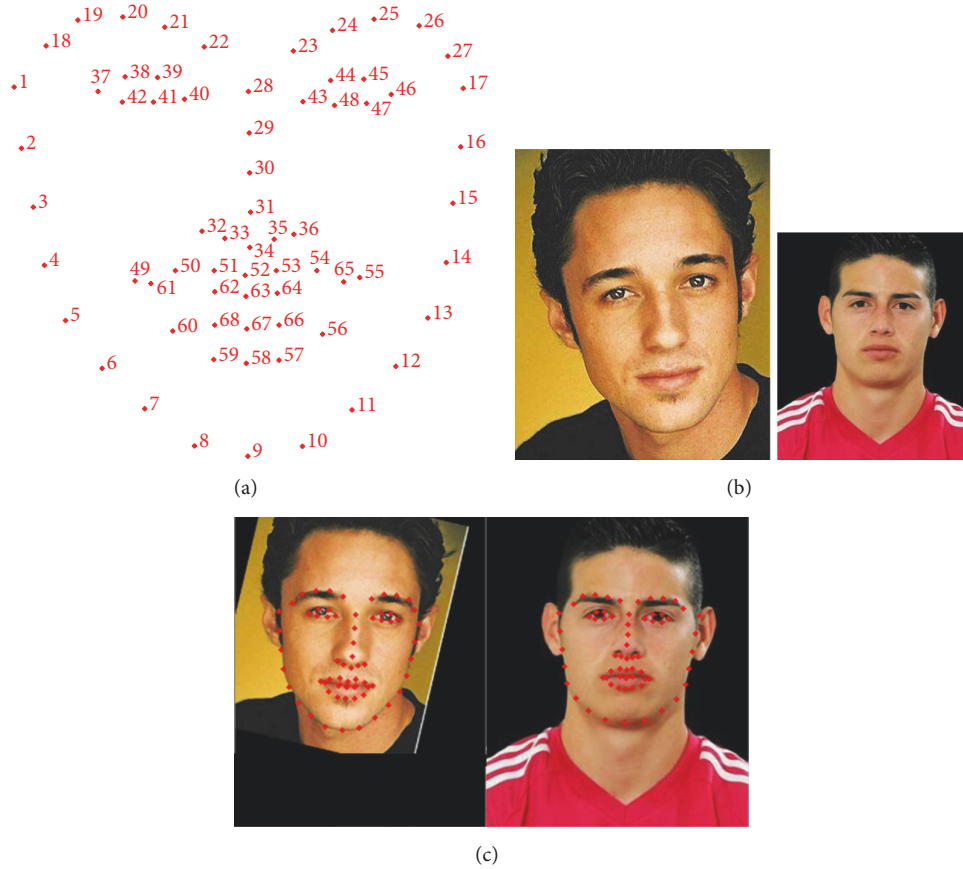


FIGURE 2: Results of facial landmark detection and face alignment: (a) definition of face landmarks, (b) input face (left) and reference face (right), and (c) results of FLD and face alignment.

the central axis  $l_s$  can be evaluated based on an optimization procedure as

$$l_s = \min_l \sum_{i \in \Omega_s} \|s_i^{(l)} - s_i\|^2, \quad (1)$$

where  $s_i^{(l)}$  is the mapping of  $s_i$  with respect to the line  $l$ . Equation (1) indicates that the central axis  $l_s$  should be the line which optimizes the symmetry of all the 68 detected facial landmarks. Consequently, the center  $c_s$  of the face is defined as the middle point of the projections of all the 68 detected facial

landmarks on the central axis  $l_s$ . The average distance from all the detected landmarks to the center point  $c_s$  is denoted by  $d_s$ , which can be used as the metric of the size of the input face. Therefore, the average distance  $d_s$  can be written as

$$d_s = \frac{1}{|\Omega_s|} \sum_{i \in \Omega_s} \|s_i - c_s\|, \quad (2)$$

where  $|\Omega_s| = 68$ . Similarly, the central axis and the size of the reference face are, respectively, denoted with  $l_r$  and  $d_r$ . Then face alignment can be implemented by rotating the input image  $I_{in}$  with the angle  $\theta$  and scaling it with the factor  $k$ , where  $\theta$  is the rotation angle from  $l_s$  to  $l_r$ , and  $k$  is the ratio between  $d_r$  and  $d_s$ , as shown in

$$\begin{aligned} \theta &= \langle l_s, l_r \rangle, \\ k &= \frac{d_r}{d_s}. \end{aligned} \quad (3)$$

After the face alignment, the original facial landmarks  $s_i$  transfer to the new locations  $s_i^a$ , and the aligned input image is denoted by  $I_{in}^a$ . Figure 2(b) shows the input face and the reference face with different sizes and poses. Figure 2(c) displays the results of facial landmark detection and face alignment. It can be seen from the result that the aligned input face is very similar in size and direction to the reference face.

**3.3. Face Warping.** To preserve the shape of the swapped face, we warp the reference image to fit the aligned input face before face replacement. The warping is implemented based on the alignment of facial landmarks. We pick 18 out of the 68 facial landmarks and denote them with  $\Phi_r = \{1, 2, \dots, 17, 34\}$  (see Figure 3(a)), which are considered to have a significant impact on facial shape. The landmarks of the reference face are denoted with  $r_i, i \in \Phi_r$ . The new locations  $r_i^w$  of most of the landmarks (except  $r_1, r_{17}$ , and  $r_{34}$ ) after the image warping should perfectly aligned to the input face, so we have

$$r_i^w = \begin{cases} r_i & \text{if } i = 1, 17, 34 \\ c_r + (s_i^a - c_s) & \text{otherwise.} \end{cases} \quad (4)$$

Figure 3(b) illustrates the original landmarks (red points) and their new locations (green points). To realize the image warping, the reference image  $I_{ref}$  is firstly decomposed into many triangle pieces based on the landmarks. The triangulation is required to minimize the change of the image background because we generally hope to preserve some parts in the background such as hair, body, and dress (that is why we do not move  $r_1, r_{17}$ , and  $r_{34}$ ). The final layout of triangulation is designed as shown in Figure 3(a). Then the image warping can be realized by applying the specific affine transformation to the corresponding triangle pieces.

Suppose that there is a triangle whose three vertices are  $r_i, r_j$ , and  $r_k$ , and the corresponding locations after the warping are denoted with  $r_i^w, r_j^w$  and  $r_k^w$ , respectively. Then the affine transformation of the triangle can be described as

$$R^w = M\hat{R}, \quad (5)$$

where

$$\begin{aligned} R^w &= [r_i^w, r_j^w, r_k^w] = \begin{bmatrix} x_i^w & x_j^w & x_k^w \\ y_i^w & y_j^w & y_k^w \end{bmatrix}, \\ \hat{R} &= [r_i, r_j, r_k] = \begin{bmatrix} x_i & x_j & x_k \\ y_i & y_j & y_k \\ 1 & 1 & 1 \end{bmatrix}. \end{aligned} \quad (6)$$

$M \in \mathbb{R}^{2 \times 3}$  is a transform matrix whose closed-form solution is  $R^w \hat{R}^{-1}$ . With the matrix  $M$ , each pixel inside the triangle can be transformed from the original location  $r$  to the new location  $r^w$ . Since the coordinates of the new locations are not integer in general, the bilinear interpolation is used to generate the final warped reference image  $I_{ref}^w$ . Figure 3(c) shows that the warped reference face has the exactly same shape as the aligned input face.

**3.4. Face Replacement.** According to the pipeline of the proposed algorithm, we need to extract the input face-ROI  $R_{in}^a$  and the reference face-ROI  $R_{ref}^w$  from the aligned input image  $I_{in}^a$  and the warped reference image  $I_{ref}^w$ , respectively. A good face-ROI should include as many facial features as possible while excluding background and distractors. However, most of traditional face swapping algorithms [2, 16] use a convex hull of the facial landmarks as the face-ROI which could cover some distracting areas like hair, hat, forehead, and neck. Therefore, in our model, a higher-precision face parsing algorithm based on deep learning [17] is introduced to extract a more accurate face-ROI. Different from the multiclass parsing network in paper [17], we use only two class labels, face and nonface, to train the parsing neural network based on the Helen dataset [18] which contains 2330 face images with pixel-level ground truth. Because our method requires face parsing rather than skin segmentation, so a contour detector is used. We initialize the encoder with pretrained VGG-16 net and the decoder with random values. During training, we fix the encoder parameters and only optimize the decoder parameters. The architecture of the network is shown in Figure 4. We set the learning rate to 0.0001 and train the network with 30 epochs with all the training images being processed each epoch. Then the face-ROIs can be generated by the retrained face parsing network.

It can be seen from Figure 5 that the retrained face parsing network can precisely extract face-ROIs in most of the cases. However, some parts of the boundary of the face-ROI  $R_{ref}^w$  obtained by the algorithm are too close to the edge of hair. If we replace the warped reference face-ROI  $R_{ref}^w$  with the aligned input face-ROI  $R_{in}^a$  directly, the boundary between the replaced face-ROI and original background of the warped reference image will become very sharp. It will disorder the results of boundary smoothing and color correction in the next step. To solve this problem, a conditional image erosion algorithm described by (7) is presented to calibrate the boundary of the  $R_{ref}^w$  and obtain a refined face-ROI  $R_{ref}^{rw}$ :

$$\begin{aligned} R_{ref}^{rw} &= \mathcal{F} \{ (x, y) \mid (x, y) \in R_{ref}^w \ \& \ SE(x, y) \cap \Omega_{edg} = \emptyset \}. \end{aligned} \quad (7)$$

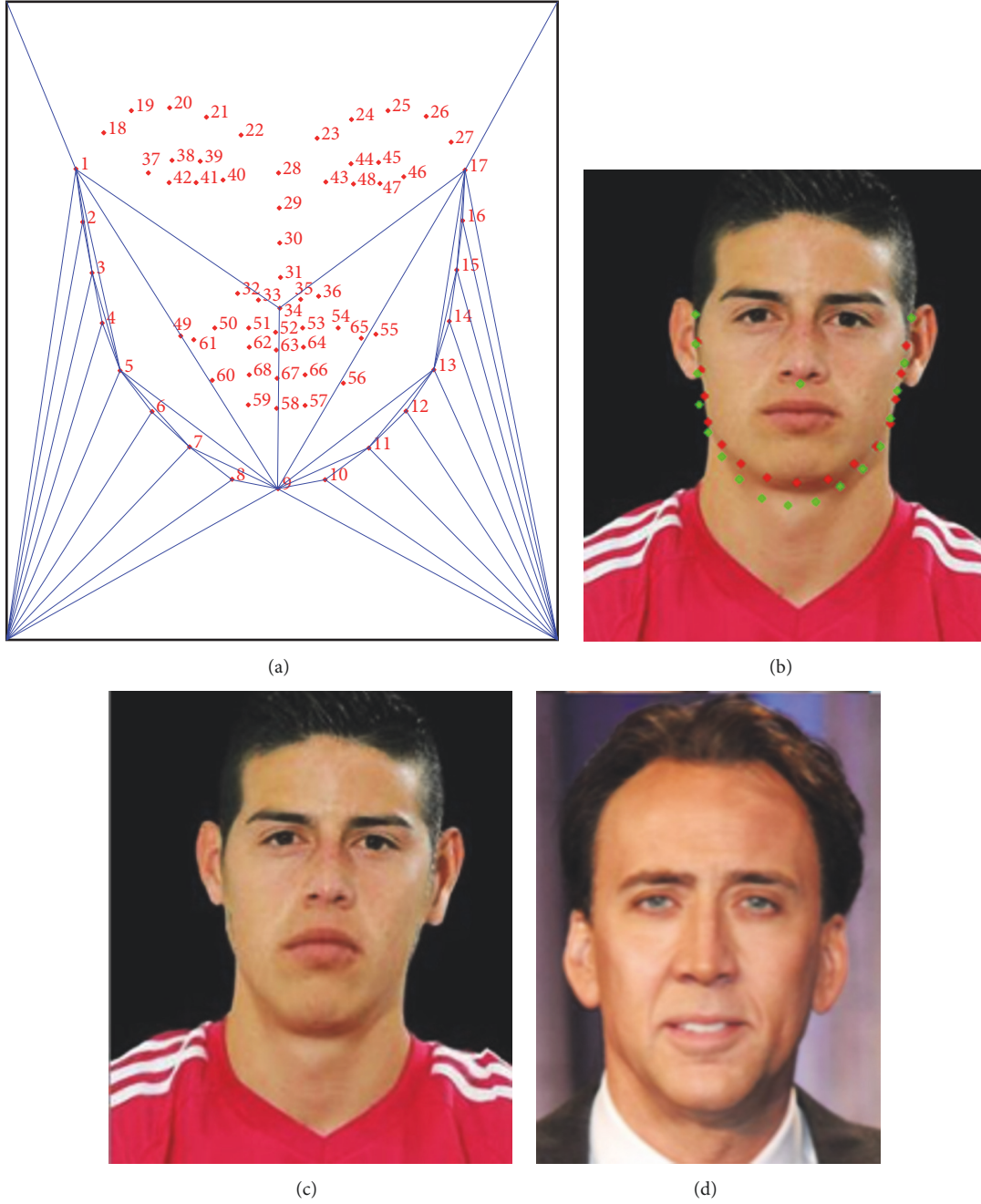


FIGURE 3: Warping of the reference face image: (a) triangulation pattern, (b) reference image  $I_{ref}$  with facial landmarks and their new locations, (c) warped reference image  $I_{ref}^w$ , and (d) the aligned input image  $I_{in}$ .

where  $SE$  is a  $5 \times 5$  square structure element and  $\Omega_{edg}$  is a set of pixels which are too close to the boundary between face and background. The exact area of  $\Omega_{edg}$  can be obtained by

$$\Omega_{edg} = \{(x, y) \mid \|g_{ref}^w(x, y)\| > g^t\} \quad (8)$$

where  $\|g_{ref}^w(x, y)\|$  is the gradient amplitude of the pixel  $(x, y)$  in the warped reference image, and  $g^t$  is a threshold of gradient amplitude. Equations (7) and (8) indicate that the

boundary of the refined face-ROI  $R_{ref}^{rw}$  can avoid the sharp edges.

Then we replace the pixels in the overlap between the refined reference face-ROI  $R_{ref}^{rw}$  and the aligned input face-ROI  $R_{in}^a$  according to

$$I_c(x, y) = \begin{cases} I_{in}^a(x, y) & \text{if } (x, y) \in R_{ref}^{rw} \cap R_{in}^a \\ I_{ref}^w(x, y) & \text{otherwise.} \end{cases} \quad (9)$$

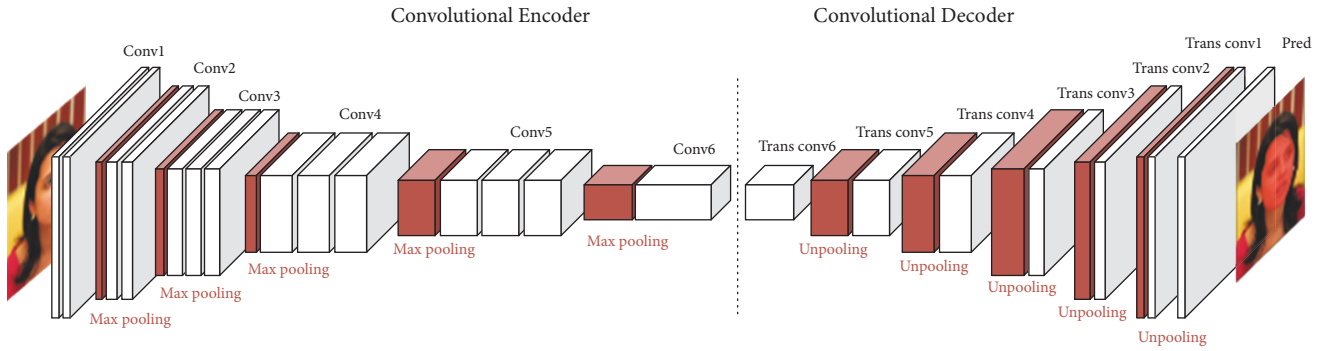


FIGURE 4: Architecture of the face parsing network.



FIGURE 5: Several typical results of human face parsing obtained by the retrained algorithm.

FIGURE 6: Illustration for face replacement and color correction: (a) warped reference face-ROI  $R_{\text{ref}}^w$ , (b) face replacement without color correction, and (c) face replacement with color correction.

Figures 6(a) and 6(b) show the refined reference face-ROI  $R_{\text{ref}}^w$  and the primary result  $I_c$  of the face replacement. It can be found from Figure 6(b) that the color of the replaced face is slightly different from the neck, ears, forehead, and other residual parts of the reference face. In order to eliminate the boundary effect and make the composite face more realistic, we use generic interpolation machinery based on solving Poisson equations [19, 20] to correct the color of the replaced face and smooth the boundary. The main idea of the color correction algorithm is minimizing the color

difference between the original pixels and the replaced ones while preserving the gradient of pixels at the boundary. It can be seen from Figure 6(c) that this method can achieve seamless clone of human face colors, making the image more realistic and natural.

#### 4. Experiment

In the experimental part, we will verify the superiority of our model by comparing it with three popular face swapping





(a)



(b)

FIGURE 7: Images used in experiment: (a) reference image and (b) input image.

TABLE 1: Similarity scores of our method and the other three methods.

| Input image   | L-model | R1     | R2     | Ours   |
|---|---------|--------|--------|--------|
|   | 89.746  | 94.85  | 87.131 | 92.549 |
|   | 87.524  | 93.727 | 90.188 | 94.06  |
|   | 89.299  | 94.201 | 89.138 | 92.864 |
|   | 85.401  | 92.975 | 85.117 | 95.202 |
|   | 87.228  | 93.896 | 75.316 | 95.256 |
|  | 84.895  | 94.83  | 81.406 | 96.012 |
|   | 86.644  | 94.711 | 86.024 | 95.017 |
|   | 85.985  | 95.334 | 78.906 | 94.41  |
|   | 91.504  | 94.008 | 76.047 | 94.027 |
|   | 85.83   | 94.649 | 77.808 | 95.927 |

algorithms [2, 7, 16] and analyze the experimental results from qualitative and quantitative aspects to explain the effectiveness of our method.

In the qualitative experiment, we use the photos of several public figures as input image and reference image, respectively, as shown in Figure 7 [7]. The five reference faces have different postures, genders, skin tones, face shapes, and hairstyles, while the input faces include a male and a female face. The corresponding visual results obtained by the competing algorithms and our model are shown in Figure 8, where the five reference faces are replaced with two input faces, respectively. Figures 8(a) and 8(b), respectively, give the face swapping results for the male input face and female input face, where each column corresponds to a reference face and each row corresponds to a face swapping algorithm. On the whole, the face swapping results of the learning-based algorithm [7] (denoted as L1) is more natural and realistic and has a good adaptability to the different perspectives, hairstyles, and skin colors. However, as mentioned above, this

method is only valid for the identities with a large number of training images, and each face identity corresponds to a generative neural network. In other words, this method cannot be applied to untrained new face images, which seriously restricts the application of this method. Our model promises the real sense of the face swapping result as well, and it is only slightly worse than the L-model in dealing with the perspective variance. In addition, this paper introduces the steps of the precise face parsing and the adaptive color correction, so it can effectively solve the problem of skin color difference and hair occlusion. While the other two replacement-based algorithms (denoted as R1 [16] and R2 [2]) have only adopted relatively simple algorithm flow, it is impossible to remove the boundary effects completely.

Many specific applications require that the facial features and face shape should remain as they should be in the face swapping, so a quantitative experiment is designed to verify the similarity between the face swapping result and the input face. Table 1 gives the similarity measures between the input



FIGURE 8: Some typical results of face swapping: (a) results of the first input image and (b) results of the second input image.

face and the swapping results shown in Figure 8, which are obtained by a CNN-based model [21]. The higher scores in Table 1 represent the higher similarity. According to Table 1, our model is obviously better than the L-model and the R2-model while slightly lower than the R1-model in some cases. This is because that the L-model produces the composite face with a slight modification of the facial features to ensure the reality of the swapping results. The R2-model does not

involve the color correction step and thus has a lower score. The R1-model has the highest similarity because it retains the complete face region at the expense of partially reducing the reality of the boundaries between forehead and hair. In contrast, our model preserves almost all of the facial features of the original input face while guaranteeing the reality of the swapped face, which leads to a better balance between similarity and realism.



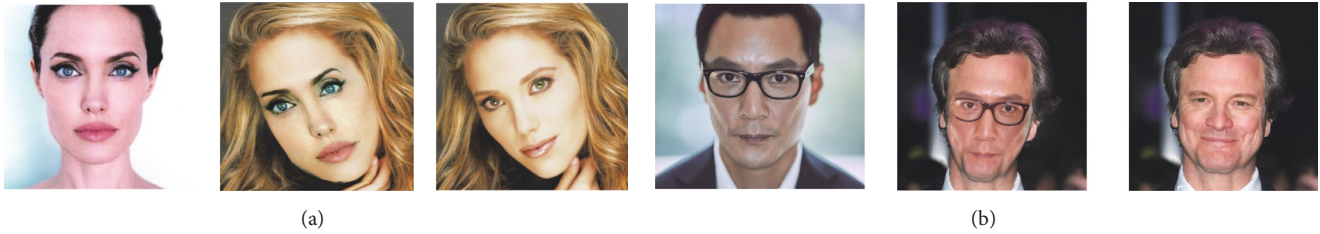


FIGURE 9: Failure examples of our method.

TABLE 2: Computational complexity and the execution time of our method and the other three methods.

|                 | L-model  | R1     | R2     | Ours   |
|-----------------|----------|--------|--------|--------|
| Time Complexity | $O(N^2)$ | $O(N)$ | $O(N)$ | $O(N)$ |
| Execution Time  | 2.32s    | 355ms  | 304ms  | 330ms  |

TABLE 3: Mean Opinion Scores of our method and the other three methods.



| Input image  | L-model | R1   | R2   | Ours |
|--|---------|------|------|------|
|   | 4.88    | 3.89 | 3.21 | 4.85 |
|  | 4.92    | 3.91 | 3.04 | 4.88 |
|  | 4.75    | 3.93 | 3.18 | 4.80 |
|  | 4.78    | 4.01 | 3.22 | 4.92 |
|  | 4.56    | 3.55 | 3.45 | 4.81 |
|  | 4.77    | 3.78 | 3.27 | 4.67 |
|  | 4.89    | 3.71 | 3.15 | 4.87 |
|  | 4.45    | 3.72 | 3.09 | 4.74 |
|  | 4.56    | 3.65 | 3.36 | 4.81 |
|  | 4.77    | 3.88 | 3.17 | 4.82 |

TABLE 4: Rating scale.

| Rating | Label     |
|--------|-----------|
| 5      | Excellent |
| 4      | Good      |
| 3      | Fair      |
| 2      | Poor      |
| 1      | Bad       |

Table 2 shows the comparison of computation complexity and the execution time of our method and the other methods. All the algorithms are implemented in Python and tested on a single core i7-8700K.

Table 3 gives the Mean Opinion Scores (MOS) of the results shown in Figure 8. The MOS is the arithmetic mean over all individual values on a predefined rating scale that a subject assigns to his opinion of the performance of a face swapping quality. It is expressed as a single rational number, typically in the range 1-5, where 1 is lowest perceived quality and 5 is the highest perceived quality. In this paper, the MOS are obtained by 100 students at Northeastern University in China according to the rating scale defined in Table 4.

Figure 9 shows some failure examples of our algorithm and the middle images are the face swapping results. The

main reason for failure shown in Figure 9(a) is that our algorithm is based on two-dimensional digital images in order to reduce time complexity, but the faces are actually three-dimensional. When the face posture difference between the input image and the reference image is too large, the result of the face swapping is unsatisfactory. Figure 9(b) shows the failure result when the face of the input image is occluded by glasses. This is mainly because the face parsing algorithm only extracts the facial features.

Figure 10 shows some other results of our method and all the images used are available in the public domain.

## 5. Conclusion

In this paper, a new face swapping algorithm based on facial landmarks detection is proposed, which can achieve fast, stable, and robust face replacement without the three-dimensional model. Our approach introduces the training results of existing learning models directly. The method we proposed does not require any training data, because it uses no learning model for new training. The experimental results show that the composite image obtained by our model has a great reality and strong adaptability to the difference of skin color and hair occlusion while retaining most of the facial features of the input image. Compared with other algorithms, our model has some advantages in aspects of visual realism,

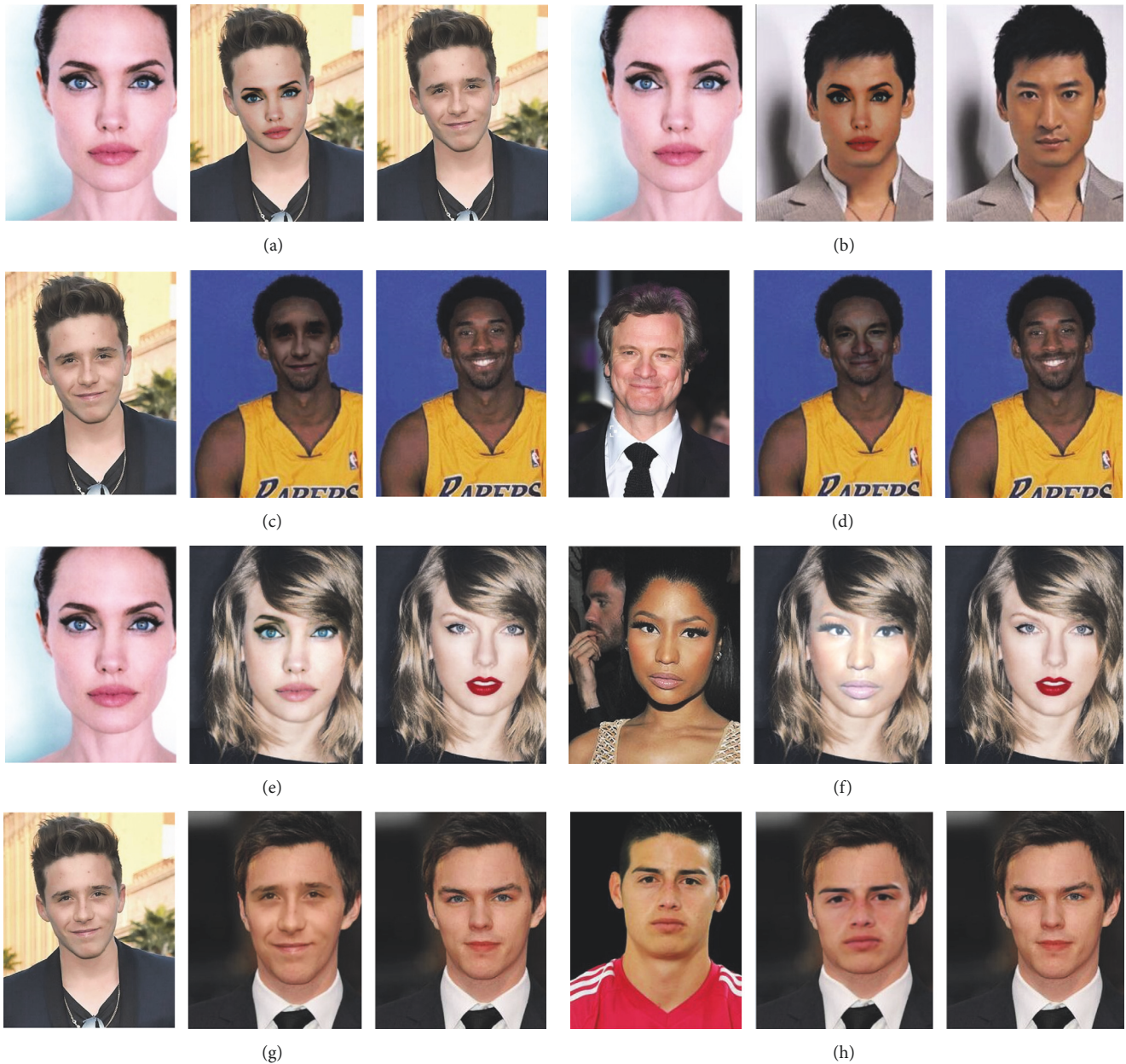


FIGURE 10: Some examples of our method.

time complexity, and data requirement. However, there is still room for further improvement, which mainly shows that the swapping result is not perfect when the input face and the reference face have a significant difference in perspective and posture. How to predict and generate the face image from a given perspective to another one is essential to solving the problem and also the main direction of our future research.

### Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

### Additional Points

*Highlights.* (i) Face alignment ensures that the input face is consistent with the size and angle of the reference face. (ii) Reference image is deformed so that the face shape of face swapping result is consistent with the input face. (iii) A face parsing algorithm was introduced to achieve better visual effects. (iv) Poisson image editing algorithm improves the realism of the results.

### Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

## Acknowledgments

This work is supported by National Natural Science Foundation of China (61701101) and by National Natural Science Foundation of China under Grant U1613214 and in part supported by the Fundamental Research Funds for the Central University of China under Grants N160413001 and N170402008, by the Fund for the Shenyang City Science and Technology Plan under Grant F16-2051-04, and by the 111 Research and Development Program of China under Grant 2016YFC0821402.

## References

- [1] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar, "Face swapping: automatically replacing faces in photographs," *ACM Transactions on Graphics*, vol. 27, no. 3, article no. 39, 2008.
- [2] S. Mahajan, L.-J. Chen, and T.-C. Tsai, "SwapItUp: A face swap application for privacy protection," in *Proceedings of the 31st IEEE International Conference on Advanced Information Networking and Applications, AINA 2017*, pp. 46–50, Taipei, Taiwan, March 2017.
- [3] Y. Nirkin, I. Masi, A. T. Tuán, T. Hassner, and G. Medioni, "On face segmentation, face swapping, and face perception," in *Proceedings of the 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*, pp. 98–105, Xi'an, China, May 2018.
- [4] V. Blanz, K. Scherbaum, T. Vetter, and H.-P. Seidel, "Exchanging faces in images," *Computer Graphics Forum*, vol. 23, no. 3, pp. 669–676, 2008.
- [5] H. X. Wang, C. H. Pan, H. F. Gong, and H. Y. Wu, "Facial image composition based on active appearance model," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2008*, pp. 893–896, Las Vegas, Nev, USA, 2008.
- [6] Y. Lin, S. Wang, Q. Lin, and F. Tang, "Face swapping under large pose variations: A 3D model based approach," in *Proceedings of the 2012 13th IEEE International Conference on Multimedia and Expo, ICME 2012*, pp. 333–338, Melbourne, Victoria, Australia, July 2012.
- [7] I. Korshunova, W. Shi, J. Dambre, and L. Theis, "Fast face-swap using convolutional neural networks," in *Proceedings of the 16th IEEE International Conference on Computer Vision, ICCV 2017*, pp. 3697–3705, Venice, Italy, October 2017.
- [8] R. Natsume, T. Yatagawa, and S. Morishima, "RSGAN: face swapping and editing using face and hair representation in latent spaces," 2018, <https://arxiv.org/abs/1804.03447>.
- [9] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014*, pp. 1867–1874, Columbus, USA, June 2014.
- [10] S. Yadav and N. Nain, "Fast face detection based on skin segmentation and facial features," in *Proceedings of the 11th International Conference on Signal-Image Technology and Internet-Based Systems, SITIS 2015*, pp. 663–668, Bangkok, Thailand, November 2015.
- [11] M. Kawulok, J. Kawulok, J. Nalepa, and B. Smolka, "Self-adaptive algorithm for segmenting skin regions," *EURASIP Journal on Advances in Signal Processing*, vol. 2014, no. 1, article no. 170, pp. 1–22, 2014.
- [12] Y. Lei, W. Yuan, H. Wang, Y. Wenhui, and W. Bo, "A skin segmentation algorithm based on stacked autoencoders," *IEEE Transactions on Multimedia*, vol. 19, no. 4, pp. 740–749, 2017.
- [13] A. Zeng, V. N. Boddeti, K. M. Kitani, and T. Kanade, "Face alignment refinement," in *Proceedings of the 2015 15th IEEE Winter Conference on Applications of Computer Vision, WACV 2015*, pp. 162–169, Waikoloa, USA, January 2015.
- [14] X. Lan, S. Zhang, P. C. Yuen, and R. Chellappa, "Learning common and feature-specific patterns: a novel multiple-sparse-representation-based tracker," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 2022–2037, 2018.
- [15] T. F. Cooles, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.
- [16] K. Dale, K. Sunkavalli, M. K. Johnson, D. Vlastic, W. Matusik, and H. Pfister, "Video face replacement," *ACM Transactions on Graphics*, vol. 30, no. 6, article no. 130, 2011.
- [17] J. Yang, B. Price, S. Cohen, H. Lee, and M.-H. Yang, "Object contour detection with a fully convolutional encoder-decoder network," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pp. 193–202, Las Vegas, USA, July 2016.
- [18] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *Proceedings of the European Conference on Computer Vision, ECCV 2012*, vol. 7574 of *Lecture Notes in Computer Science*, pp. 679–692, Springer, Berlin, Germany, 2012.
- [19] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," *ACM Transactions on Graphics*, vol. 22, no. 3, pp. 313–318, 2003.
- [20] J. M. Di Martino, G. Facciolo, and E. Meinhardt-Llopis, "Poisson image editing," *Image Processing On Line*, vol. 5, pp. 300–325, 2016.
- [21] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," 2015, <https://arxiv.org/abs/1504.03641>.

