IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 May 2024, accepted 15 June 2024, date of publication 21 June 2024, date of current version 28 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3417442

RESEARCH ARTICLE

Conditional-GAN-Based Face Inpainting Approaches With Symmetry and View-Degree Utilization

TZUNG-PEI HONG[1,2], (Senior Member, IEEE), JIN-HANG WU[1], JA-HWUNG SU[1], AND TANG-KAI YIN[1]

[1]Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung 811726, Taiwan
[2]Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung 804201, Taiwan

Corresponding author: Ja-Hwung Su (bb0820@ms22.hinet.net)

This work was supported by the National Science and Technology Council, Taiwan, R.O.C., under Grant NSTC 113-2622-E-390-002 and Grant NSTC 112-2622-E-390-004.

ABSTRACT Recently, image inpainting has been proposed as a solution for restoring the polluted image in the field of computer vision. Further, face inpainting is a subfield of image inpainting, which refers to a set of image editing algorithms re-conducting the missing regions in face smoothly. Actually, face inpainting is more challenging than general image inpainting because it needs more face structure information. Although a number of past studies were proposed for face inpainting by using face segmentation, face edge and face topology, there is some important information ignored, such as geometric and symmetric properties. Based on such concepts, in this paper, we propose a two-stage face inpainting method called CGAN (Conditional Generative Adversarial Network) which integrates face landmarks and Generative Adversarial Network (called GAN). In the first stage, the face landmark is predicted as the condition, providing GAN with important information of geometry and symmetry. The main idea in this stage is to dynamically adjust the loss by the proposed view degree. Accordingly, the masked face image and the corresponding face landmark are used as conditions input to the GAN in the second stage. Finally, the missing-regions are inpainted by the proposed CGAN. To reveal the effectiveness of proposed method, a number of evaluations were conducted on real datasets. The experimental results show that, the proposed method predicts a better face landmark by information of geometric structures and symmetric outlooks, and thereupon the proposed CGAN reconstructs the missing regions superior to the compared methods.

INDEX TERMS Face inpainting, face-landmark, generative adversarial networks, deep learning, autoencoder.

I. INTRODUCTION

The purpose of image inpainting—also known as image completion and image hole filling—is to reconstruct or fill missing areas of an image with natural and plausible contents. Image completion can also be regarded as image synthesis, in which the perception of the synthetic images should be as realistic as possible. In recent years, given the advances in hardware, computing power has dramatically increased in general, enabling the active development of deep learning (DL) related technologies. Thus many DL-based image processing technologies have been proposed. Likewise, DL-based image inpainting technology now processes images with complex textures. Moreover, as neural networks deepen, they can gradually consider semantic-level information to help complete damaged images. In the field of image inpainting, face inpainting is one of applications. The main intent of face inpainting is to achieve identifications of the photos in ID cards and images in surveillance videos, as shown in Figure 1. For example, if the visual face in a surveillance video is not clear, you can restore the image by face inpainting. Yet, it is more challenging than general image inpainting because of the following difficulties.

In terms of the occlusion area size, assume that the occluded face areas are large in an image. For a successful inpainting, the face features should be restored in the correct

The associate editor coordinating the review of this manuscript and approving it for publication was Valentina E. Balas.

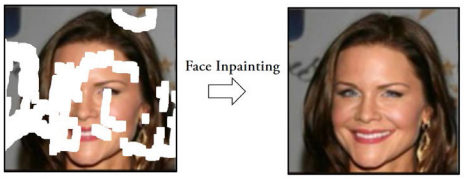

FIGURE 1. Scenario of face inpainting.

positions. On the contrary, facial features generated in the wrong place will result in unnatural face images. For example, the eyes generated under the nose or the nose positioned to the left of the mouth. Therefore, a successfully restored face might be symmetric and have consistent attributes even when only a small area is occluded. Facial features in the filled area that differ from other facial features likewise result in strange-looking faces. For instance, it would be odd if the original image had a larger right eye shape with makeup, but the left eye was inpainted with a smaller eye shape and no makeup.

Moreover, for the generated content, noticeable seams between the filled content and the other content in the face image make the image artificially synthesized and unnatural. Regarding the person's pose in an image, as the person does not always face the camera due to different shooting angles, the face is sideways in some images. When the person's posture in a two-dimensional image is sideways, the sizes of the left and right halves of the face differ greatly, as do the proportions of similar facial features. This further complicates image inpainting.

To address the above problems, in this paper, we propose a two-stage method extended from the past study [11], achieving high quality of face restoration by integrating face landmark and GAN. Also, it can be called Conditional GAN because the face landmark can be viewed as a condition for GAN. In the first stage, we use the face datasets to train a prediction model for face landmarks. We also use a weighted loss function to improve the model ability to predict face landmarks with complex facial feature locations, such as a side-view face. In the second stage, we perform symmetry processing to generate reasonable contents with which to fill the missing areas. Finally, the inpainted result is post-processed to reduce noise and artifacts. In overall, the contributions can be folded as follows.

- A two-stage training method is proposed to train a deep learning-based model to achieve the face inpainting. The completed image produced by our approach can not only maintain the geometric structure but also keep the attributes consistent.
- Although the landmark has been studied in recent face inpainting, the view degree is ignored. In this paper, the face view-degree is calculated first, and a weighted loss function is executed accordingly. It effectively improves the prediction result of face-landmark, increasing the robustness of the landmark prediction model and alleviating the problem of data imbalance.
- In addition to view-degree, the other contribution is the symmetry processing proposed to improve the symmetrical outlook of the inpainting result, which makes the synthetic content rational.

To realize the effectiveness of proposed method, a set of evaluations were made on real datasets, referring to face landmark prediction and face inpainting. For face landmark prediction, the proposed method with geometric and symmetric information predicts a better face landmark than the compared methods. Further, the proposed face inpainting based on the face landmark reconstructs the missing face regions more successfully than the competitor. The rest of this paper is structured as follows. In the second section, the related works are briefly reviewed. The proposed method is presented in Section III. In Section IV, the related experimental results are analyzed. Finally, the conclusion and future works are given in Section V.

II. RELATED WORK

On the whole, the related works can be classified into three categories, namely traditional image inpainting, autoencoder and generative adversarial network, which are described in the succeeding in this section.

A. TRADITIONAL IMAGE INPAINTING

In the era when deep learning was not yet widespread, many traditional image inpainting methods used texture synthesis technology to generate images by filling in the missing areas, such as the non-parametric method proposed by Efros and Leung [8]. It will form a non-parameter probability distribution function of all similar neighborhoods and finally calculate the similarity between the source area and the aimed area via selecting the most similar area. This method filled images with pixels, so the speed is very slow [8]. In order to tackle this problem, the same authors proposed an improvement in two years and changed the pixel to block as the filling unit, which has greatly improved the speed [7]. Criminisi et al. proposed an algorithm to remove large objects in digital images. The method first calculates the priority of the points to be filled according to the confidence value of the pixels and image isophotes, then searches for the closest block in the known area, copies it to fill in the target area, and finishes the image inpainting task [4]. Barnes et al. proposed the patch-match algorithm using the image continuity to reduce the search range greatly and to ensure that most points can converge fast by iterating, reducing the computational complexity and speeding up image completion [1]. Wang et al. also proposed an algorithm for repairing texture images using texture synthesis. This method decomposes texture images into cartoon and texture images and restores the structure of the missing areas of the image based on the boundary. Although their approach can solve most boundary problems in image inpainting, the generated image is still slightly blurred, and the effectiveness is not very good [37]. In summary, although the methods mentioned above all have an algorithmic solution to the specific problem, they still have

a common disadvantage. First, they can only fill unknown areas according to the information of the existing area and cannot predict and fill in what is not present in known areas. Second, their application in known regions is limited mainly to parallel movement. If variations such as sizes and rotations are added, the overall recovery performance degrades due to the increased computational complexity. Besides above, Dang and Lee [6] attempted to achieve high quality of scene text segmentation. To this end, it used CGAN-based image inpainting to synthesize robust scene text images. The aim of this related work is different from ours, so that the condition of GAN is different. Richardson et al. [28] proposed an image inpainting method named pixel2style2pixel (pSp), consisting of an encoder with a feature pyramid and multiple mapping networks. Then the pre-trained StyleGAN generator is performed as conditions to restore the occlusion image.

B. AUTOENCODER

The autoencoder (AE) was first proposed by Rumelhart et al. [29]. Its architecture consists of an encoder and a decoder. The encoder reduces the image dimensions and extracts the compressed low-dimensional features which can also be regarded as an essential image feature. Conversely, the decoder decompresses this low-dimensional representation to a high-dimensional image which can be viewed a reconstructed operation according to the crucial features of the original image. For autoencoder, a low-dimensional representation is obtained after inputting the original image into the encoder. The obtained result is then input to the decoder to reconstruct the image. The network is trained using loss functions and backpropagation to produce a reconstructed image as similar to the original input image as possible. Thus the main idea of AE is to learn the feature representations of the input data in an unsupervised manner. Nevertheless, AE can only be decoded into specific data through a particular low-dimensional representation, and its ability to generate diverse samples is limited. With the rapid growth of deep learning technology, various types of autoencoders have been proposed, including denoising autoencoders [36], contractive autoencoders [30], and variational autoencoders (VAE) [16]. Among these, the variational autoencoders proposed by Kingma et al. became quite popular. VAE [16], an improved version of AE [29], enhances the ability to extract features. It added noises to the feature vector generated by the encoder and considered the characteristics of the normal distribution to make the generated feature vectors more diverse. Therefore, it is suitable for data generation. Many recent image generation methods are based on VAE [16]. Cai et al. proposed a VAE-based image synthesis method combining residual network concepts and skip connections [12] to optimize the original VAE model [3]. They adopted a coarse-to-fine multi-stage approach in their network architecture to generate higher-resolution images than VAE [16]. Oord et al. proposed the vector quantized variational autoencoder (VQ-VAE) [35], an improved version of VAE [16] that leveraged vector quantization to prevent posterior collapse and generated high-quality images. Some methods for face generation are also based on VAE [16]. For example, Qian et al. proposed the additive focal variational autoencoder (AF-VAE) to generate faces with different expressions. Its architecture is likewise an improved version of VAE [16], which combines face structure and appearance information for face modeling to achieve good expression synthesis [27]. Tu and Chen proposed a VAE-based face generation model searching for possible VAE encoding vector sets for occluded images and restores the face appearance using the decoder [34]. Various image completion methods based on GAN [10] have been proposed. Liu et al. proposed partial convolution for image completion, filtering out valid pixels in an area and then renormalizes them to generate an image [18]. Liu et al. proposed a novel coherent semantic attention layer for image completion which acquires contextual information near the missing areas of the image so that the generated images maintain semantic consistency [19]. Other GAN-based image completion methods include PEN-Net [39] and 3DFaceFill [5]. Likewise, many face inpainting methods are based on GAN [10] or CGAN [23]. For example, Liu et al. proposed a generative face-inpainting model [23] and trained by combining reconstruction loss and semantic parsing loss to generate part of the face in missing areas [17]. Cai et al. proposed the face inpainting and face super-resolution generative neural network (FCSR-GAN), which learns a model based on multi-task learning that complements the face and improves the face resolution [2]. To train their network to yield good generation effects, they used adversarial loss, style loss, and smooth loss functions. Yang et al. proposed a generative landmark-guided face inpainting method (LaFIn), using face landmarks corresponding to facial features to guide face generation [38]. In the proposed method, we leverage LaFIn [38] and further improve the shortcomings of recent face inpainting methods.

C. GENERATIVE ADVERSARIAL NETWORK

The generative adversarial network (GAN) was first proposed by Goodfellow et al. [10]. The GAN architecture mainly comprises a discriminator and a generator. The discriminator discriminates the true degree of the results generated by the generator as accurately as possible, whereas the generator yields results that closely resemble real data that the discriminator cannot correctly discriminate its true degree. The central idea is to strengthen the ability of generating data via competition between the two networks. After GAN was proposed, numerous extensions appeared, including conditional-GAN (CGAN) [23], CycleGAN [41], and StyleGAN [15]. One commonly used architecture is conditional-GAN (CGAN) [23]. The difference between CGAN [22] and GAN [10] is that the former [23] generates images based on the input label as a condition that specifies the generated result. Given this condition, the generator is trained to generate results that are more in line with the expected results. Recently, several well-known applications

FIGURE 2. Face landmark [31].

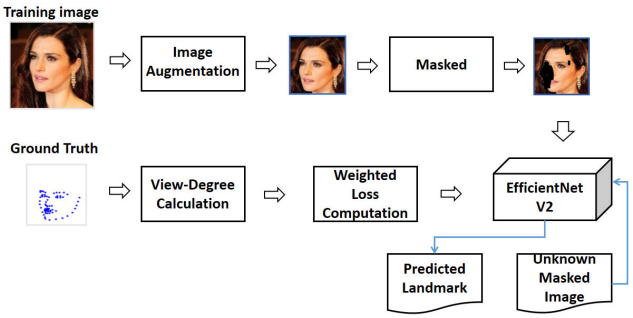

FIGURE 3. Workflow of face landmark prediction.

apply CGAN [23] as the basic architecture. For example, Isola et al. proposed an image conversion method that uses the loss function to learn a mapping between the conditional input image and the ground-truth image, generating an image that reflects the input condition [12]. Applications such as text and image conversion [40], video and audio generation [22], and image editing [26] have also been developed with this concept. PatchGAN proposed by the authors of pix2pix [13] changes the discriminator to a fully convolutional layer so as to improve discrimination. It works by discriminating the "realness" of each local image patch to determine whether the entire image is real. In contrast to GAN, the patch-GAN discriminator [13] is different from the general GAN discriminator [10]. In patch-based GAN, the discriminator discriminates the true degrees of all N^*N patches of the input image, resulting in N^*N scores p_i. After summing all p_i values, they are averaged into p, which is used to represent the true degrees of the overall image.

III. PROPOSED MATHOD

Although a number of previous works were proposed on image inpainting, the effectiveness for face inpainting can be improved by considering the geometric and symmetric properties of a face. To achieve this idea, in this paper, we propose a two-stage face inpainting method, including face-landmark prediction and GAN-based inpainting. For face-landmark, it can be viewed as a condition supporting the GAN, while GAN is a generative component to restore the incomplete image face. In this section, the proposed face inpainting method will be presented in detail.

A. FACE-LANDMARK PREDICTION

Face landmark is a set of points representing the positions and contours of essential organs of the face. In this paper, the face landmark is composed of 64 points, namely the outline of the face (1–17), the right eyebrow (18–22), the left eyebrow (23–27), the nose (28–36), the right eye (37–42), the left eye (43–48), and the mouth (49–68), as shown in Figure 2.

In traditional face inpainting, the large masked area will lose much face feature information, making the inpainting un-robust. To deal with this issue, the face landmark is used to provide GAN with effective location information of the aimed region instead of the whole image. Moreover, other important face landmark information such as the shape and size of facial features can also be provided to GAN. Therefore, the inpainting results will be better. Figure 3 shows the workflow of face landmark prediction, including the training and testing phases. In the training phase, we input the general face image to the image augmentation module and randomly augment images with different data augmentation methods. The purpose of this data augmentation is to increase the diversity of the training dataset so that the training model can better predict face at any position in the image. Since the size and position of the face may change when the image is augmented, we adjust the ground-truth face landmark obtained from the training dataset to match the augmented face. Then, we use the view-degree calculation module to calculate the view-degree of the face in the input image based on the ground-truth face landmark and generate a weight value λ with which we adjust the penalty value of the loss function to improve the ability to predict images with side faces.

At the same time, to train the model predicting the various positions of face landmarks even in masked face images, we randomly add masks to the augmented images and then input them to the face-landmark prediction model for training. Because EfficientNetV2 [33] is smaller in scale and faster to train than other models with the same high accuracy and is thus suitable for our task and dataset, we chose it as the backbone of the model. We set the output size of the last fully connected layer to 136, corresponding to the x and y coordinates of the 68 facial key points. Subsequently, we input the predicted face landmark and ground-truth face landmark to the loss function, obtaining the loss value between the actual and predicted values. We multiply the loss value by λ so that it changes with the magnitude of the view-degree of the face. In this way, the model is trained to improve its robustness for side images. In the testing phase, after inputting the masked face image to the face-landmark prediction model, the face landmark is predicted from the model.

1) IMAGE AUGMENTATION

As mentioned above, the training data is randomly augmented with an image augmentation module before training the face-landmark prediction model. In this module, we use five data augmentation modes to process images, namely origin mode, alignment mode, enlargement mode, reduction mode, and noise mode. For the origin mode, it directly uses the

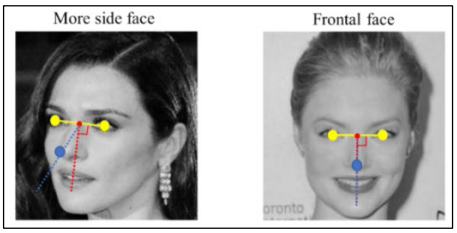

FIGURE 4. Scenario of view-degree judgment.

TABLE 1. Notations of parameters for view-degree calculation.

Symbol	Notation
I	The image in the training dataset
I^a	The image I after augmentation
I^m	The image I^a after adding a mask
L^{gt}	The ground-truth face landmark of image I
L^p	The predicted face landmark of image I
θ	The view-degree of the face in the image
e^l	The center of the left eye
e^r	The center of the right eye
e^c	The center of the eyes
n^c	The center of the nose
v^e	The vector between the eyes
v^v	The vector perpendicular to the vector $\vec{v}$
v^f	The vector presenting the view-degree of the face in the image
R_θ	The list stores the normalized values from N_θ
λ	The weight to adjust the loss penalty between L^{gt} and L^p
α	The weight to λ range

original image for training, and the alignment mode aligns the face in the image so that the eyes are vertically centered within the image. In the enlargement mode, the proportion of the area containing the face is enlarged, and in the reduction mode, the image is shrunk, and pixels randomly padded around it to the required size. The purpose of these two modes is to train the model for better prediction results given images with large or small face proportions. In the noise mode, we randomly generate salt and pepper noises at the pixel level, changing the pixel color to white as ''salt'' noises, and changing them to black as ''pepper'' noises. With this mode, we seek to train the model for better prediction results on noisy images.

2) VIEW-DEGREE CALCULATION

The goal of view-degree to adjust the loss function to improve face-landmark prediction. As shown in Figure 4, a frontal face indicates the center of the nose lies just below the middle of the eyes, while a more side face indicates the area of one half of the nose is larger than the other half.

Based on Table 1, the algorithm for view-degree calculation is listed as follows, where the input is the ground-truth face landmark L^{gt} of an image.

Step 1: Sum coordinates 43–48 in the landmark L^{gt} representing left eye contour and average to obtain the center of the left eye e^l.

Step 2: Sum coordinates 37–42 in the landmark L^{gt} representing right eye contour and average to obtain the center of the right eye e^r.

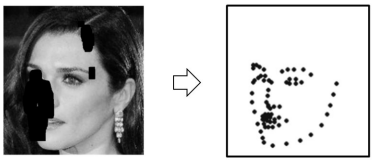

FIGURE 5. Resulting examples of face landmark prediction.

Step 3: Calculate center coordinates of eyes e^c as follows:

$$e^c = (e^l + e^r)/2$$

Step 4: Calculate vector v^e $(ve\ x,\ ve\ y)$ between the eyes of the face image and calculate the vector v^v $(vv\ x,\ vv\ y)$ perpendicular to the vector v^e as follows:

$$v^e = e^r - e^l, \text{ and } v^v = (-v_y,\ v_x)$$

Step 5: Use coordinate 31 in the landmark representing the center of nose n to calculate the vector v^f $(vf\,x,\ vf\,y)$ showing the view-degree of the face as follows:

$$v^f = n - -e^c$$

Step 6: Use the Cosine Theorem to calculate the view-degree θ of the face as follows:

$$\boldsymbol{\theta = cos^{-1}} \frac{\boldsymbol{v^f \cdot v^v}}{|\boldsymbol{v^f}| \times |\boldsymbol{v^v}|}. \tag{1}$$

Step 7: Return View-degree θ of the face in the image.

3) WEIGHTED LOSS COMPUTATION

In the land mark prediction stage, the proposed loss function can be viewed as an improved L2 norm loss function, which is shown in Equation (2).

$$Loss\left(y, y'\right) = \lambda \left(\sum_{j=0}^{n} \left|y_j - y'_j\right|^2\right)^{\frac{1}{2}}, \tag{2}$$

where y represents the ground-truth landmarks, y' represents the predicted landmarks, and y_j represents the coordinates of the j-th point of y. Here, λ is a parameter used to adjust the loss penalty to penalize images with a larger view-degree, which is defined in Equation (3).

$$\lambda = 1 + (sin\theta)\,(1 - R_\theta)\,/\alpha. \tag{3}$$

where θ represents the view-degree of the image, R_θ represents the normalized number of images for θ and weight α is used to adjust the range of λ. For example, assume that the maximum and minimum numbers of images in the training set are 8806 and 10, respectively. If the number of images for θ =30 is 1709, the R_θ will be (1709-10)/(8806-10)≈0.1932. The basic idea behind R_θ is that, the amount of data available for each view-degree highly affects model training. This idea increases the sensitivity of networks trained to predict landmarks of such faces, and has a greater influence on updating the weights. That is, with λ, the ability of the model to predict the image with a side face can be enhanced.

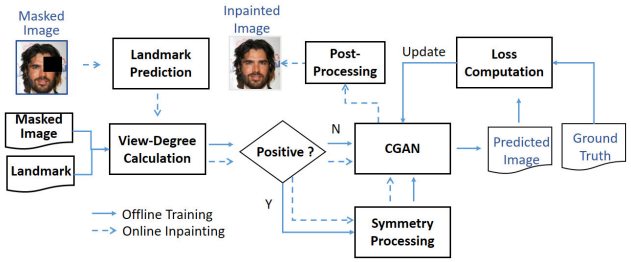

FIGURE 6. Workflow of face inpainting.

4) FCAE LANDMARK PREDICTION

In this paper, EfficientNetV2 is used as the face landmark prediction model. Based on Equation (2), the parameters of the face-landmark prediction model will be updated well. For an unknown image, through EfficientNetV2, the masked face can be landmarked. Figure 5 is an illustrative example for face landmark prediction results.

B. GAN-BASED FACE INPAINTING

In this stage, we concatenate the predicted landmark with the masked face image and input it into the face-inpainting model to complete all parts of the face and fill in the missing areas with reasonable and less strange content. As shown in Figure 6, the face-inpainting stage is mainly divided into two phases, namely offline training and online inpainting phases.

In the training phase, first, the view-degree of a training image is calculated as a view-degreed image according to the facial features information provided by the face landmark. Next, the view-degreed image is determined as positive or negative. If negative, CGAN is trained directly by the view-degreed image. If positive, the symmetry processing is performed to obtain the processed image according to its view-degree. In addition, we input the processed image into the CGAN model to complete and restore the missing area in the image. Finally, the image without occlusion generated by the neural network is output. To make the final CGAN model better, we input the generated image and the original ground-truth image into the loss function to calculate the loss, and continuously train the model to reduce the loss value. In the inpainting phase, for a masked image, the face-landmark is predicted first. Furthermore, the view degree is calculated to determine if performing symmetry processing. Next, the CGAN model fills in the missing areas of the face in the processed image. Finally, the inpainted region is smoothed by post-processing. In the following subsections, the symmetry processing, CGAN and post-processing will be presented in detail.

1) DETERMINATION OF POSITIVES AND NEGATIVES

This operation is mainly used to determine the processed image by view-degree θ. If $\theta \leq 20$, the view-degreed image is identified as positive, otherwise, negative. The threshold used to determine the view-degree class is set to be 20 because we have observed that when the view-degree of the face is lower than 20, the facial features of the left and right faces still remain the left-to-right symmetry. Therefore, we regard the face with a view degree below 20 as a nearly-frontal face.

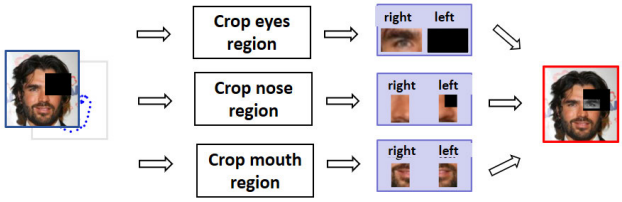

FIGURE 7. Example of symmetry processing.

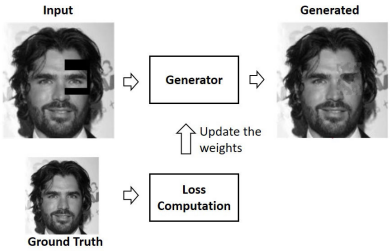

FIGURE 8. Example of generator of CGAN.

2) SYMMETRY PROCESSING

The images belonging to the positive class tend to have a more frontal face, so we will perform symmetry processing on the areas of the located eyes, nose, and mouth. First, the regions of the three organs in the face are cropped by extending the most up, down, left and right margins to bigger margins. Second, the occlusion ratio for each organ is calculated, which can be defined as: O_a (size of occluded area / size of the cropped area). Third, If the occlusion ratio is larger than 70%, the symmetry part for each organ such as left-to-right or right-to-left is copied and flipped. Finally, the flipped ones are colored into a grayscale type and then pasted to the occlusion part. Figure 7 is an example for symmetry processing. In this example, because the occlusion ratios of the nose and mouth do not exceed the occlusion threshold 70%, the right eye is just flipped and copied to the left eye in a grayscale format.

3) FACE INPAINTING BY CGAN

By referring to Figure 6, whatever for positives or negatives, the CGAN is trained in the offline phase, consisting of a generator and a discriminator. So-called Conditional GAN indicates the generative network considering conditions of the geometric structure and the attributes consistent. In the proposed method, Attention U-Net [25] is used as the backbone of the generator. The concept of Attention U-Net [25] is that in the second half of the entire network, an attention gate would be added to obtain an attention vector representing the importance of different regions in the feature map and be added to the output result of each layer for decoding. It is related to the effectiveness of our face inpainting task because the model can be used to learn the characteristics of essential regions to help complete the face. Figure 8 is an example of generator of CGAN. In this example, as we

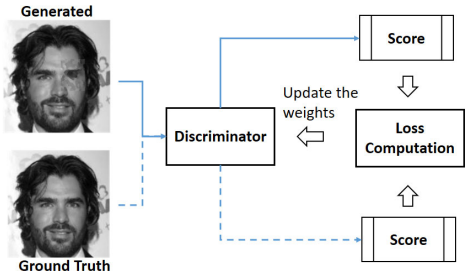

FIGURE 9. Workflow of discriminator of CGAN [13].

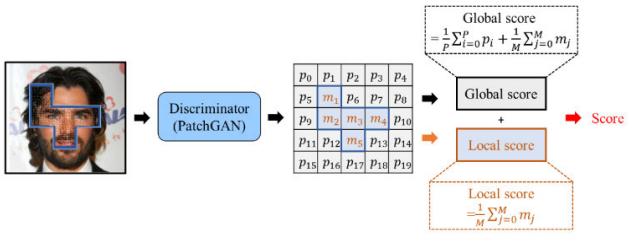

FIGURE 10. Example of generating the authenticity score for the discriminator of CGAN [13].

can recall from the contribution of symmetry processing, the input is a positive image symmetrically processed. Through the iteratively updating the weight, the CGAN will be better and better.

For generator, the loss is calculated by a hybrid loss function integrating Adversarial Loss [21], Perceptual Loss [14], Style Loss [9], Pixel Loss, Total Variation Loss [24] and Prior Loss [16], which is defined as:

$$L_G = w_0 L_{adv} + w_1 L_{per} + w_2 L_{style} + w_3 L_{pixel} + w_4 L_{tv} + w_5 L_{prior}, \tag{4}$$

where w_0, w_1, w_2, w_3, w_4 and w_5 are the weights set by referring to LaFIn [38], L_{adv} is Adversarial Loss, L_{per} is Perceptual Loss, L_{style} is Style Loss, L_{tv} is Total Variation Loss, L_{prior} is Prior Loss, and L_{pixel} is Pixel Loss. Here, the Pixel Loss aims to calculate the difference between pixels of the real image and the generated image. The equation of it is as follows:

$$L_{pixel} = \frac{1}{n} \sum_{i=1}^{n} |y_i - x_i|, \tag{5}$$

where x_i is the i-th pixel in the generated image, y_i is the i-th pixel in the real image, and n is the number of pixels in the image.

For the discriminator, patchGAN [13] is used to determine whether the whole image is real or not by discriminating the authenticity of each local image patch. As shown in Figure 9, the discriminator is optimized by iteratively updating the weights based on the loss where the loss is calculated by the authenticity score. Since not all areas of the face image are occluded, after the discriminator outputs the value of all patches, we calculate two kinds of scores regarding the authenticity of the image, namely global score and local score. The global score is the average of the scores of all patches, which is defined as:

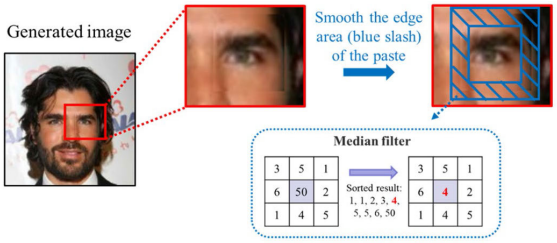

FIGURE 11. Example of noise smoothing.

TABLE 2. Experimental datasets.

Dataset	Title	#Images
Face	CelebA	202,599
Datasets	300W	3,837

$$\frac{1}{P} \sum_{i=0}^{P} p_i + \frac{1}{M} \sum_{j=0}^{M} m_j \tag{6}$$

where P indicates the size of non-patches, M indicates the size of patches, p_i indicates the i^{th} pixel score surrounding the patch and m_j indicates the j^{th} pixel score in the patches. The local score is the average of the scores of the occluded patches, which is defined as:

$$\frac{1}{M} \sum_{j=0}^{M} m_j. \tag{7}$$

Then the two scores are added to obtain the authenticity score of the entire image. Figure 10 depicts the example of generating the authenticity score for the discriminator of CGAN.

Since the purpose of the discriminator is to judge whether the generated image is real or not, we select the adversarial loss to calculate the loss of the realness of the generated image and the real image. The equation for adversarial loss is as follows:

$$L_D = \frac{1}{n} \sum_{i=1}^{n} \frac{(1\text{-D}(y))_i^2 + D(G(z))_i^2}{2}. \tag{8}$$

where z is the input masked image, D is the discriminator, G is denoted as the generator, $G(z)$ is denoted as the generated image, $D(G(z))$ is the result vector of the generated image after being discriminated, and n is denoted as the number of dimension in the vector output from the discriminator. The main objective of this loss function is to train the discriminator to identify real or fake images correctly. This loss function comes from Least Squares Generative Adversarial Networks (LSGANs) [21], which is not the same as that of the discriminator of the traditional GAN. The traditional GAN uses the cross-entropy loss, and however, LSGANs [21] improves the loss calculation method, using the least square method to calculate the loss, making the training model more stable and better.

4) POST-PROCESSING

In the online stage, after CGAN, the generated image would be post-processed to eliminate some noises caused by symmetry processing and to improve the visual effectiveness of the generated image. In this stage, if there is a symmetrically processed area, its position parameters will be recorded and sent for noise smoothing. As shown in Figure 11, the pixels would be smoothed in the area around the seam (the blue slashed area) by the median filter, reducing the artifacts and noises caused by pasting. The median filter will sort all pixel values in the area to be filtered and then obtain the median to replace the pixel value in the center of the area.

IV. EMPIRICAL STUDY

A. EXPERIMENTAL DATA

In this paper, the experimental data consists of two sets, namely CelebA [20] and 300W [31], [32]. We apply an 8:1:1 split of CelebA dataset for training, validation, and testing, respectively. The training set is used to train the model for optimizing parameters, the validation set is used to obtain the best model in training, and the testing set is used to evaluate the performance of the trained model. Details of these datasets are introduced in Table 2.

B. EVALUATION MEASURES

In the experiment, we use the measures mentioned in [31], namely inter-ocular normalization (ION) and inter-pupil normalization (IPN), measuring the performance of our face-landmark prediction models trained with different parameters. For face inpainting, we use peak signal-to-noise ratio (PSNR) to measure the degree of distortion of the image after inpainting. We also use structural similarity (SSIM) to measure how similar the completed and original images are. These measures are described separately below.

1) INTER-OCULAR NORMALIZATION

Inter-ocular normalization (ION) calculates the loss between the predicted and ground-truth landmark [31]. It normalizes the loss with the distance between the outer corners of the two eyes to eliminate unreasonable changes caused by non-uniform face scales. The ION is defined as:

$$ION\left(y, y'\right) = \sum_{i=1}^{N} \frac{\left|\left|y_i - y_i'\right|\right|_2}{d_{ion}}, \tag{9}$$

where y is the ground-truth landmark, y' is the predicted landmark, y_i is the i-th key point in y, y_i' is the i-th key point in y', N is the number of landmark points, and d_{ion} is the distance between the outer corners of the eyes.

2) INTER-PUPIL NORMALIZATION

Inter-pupil normalization (IPN) is very similar to the equation (15), and they only differ in how the distance used for normalization is calculated [31]. The IPN is defined as follows:

$$IPN\left(y, y'\right) = \sum_{i=1}^{N} \frac{\left|\left|y_i - y_i'\right|\right|_2}{d_{ipn}}, \tag{10}$$

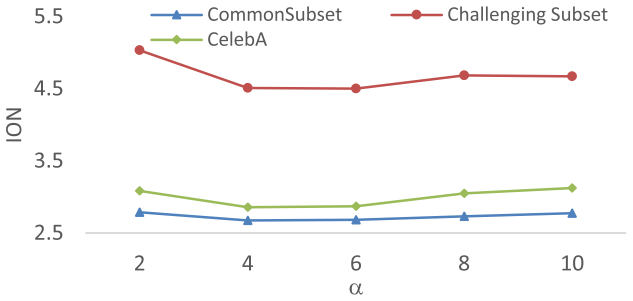

FIGURE 12. IONs for different α settings, datasets and measures.

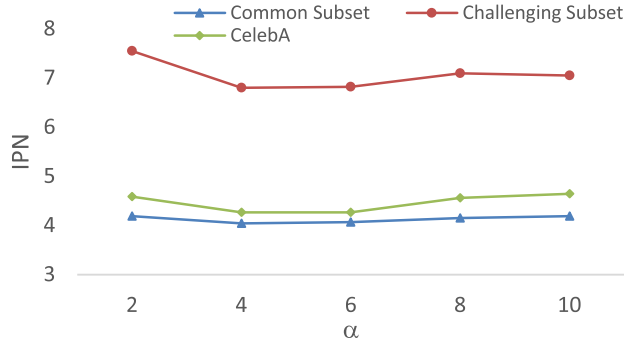

FIGURE 13. IPNs for different α settings and datasets.

where y is the ground-truth landmark, y' is the predicted landmark, y_i is the i-th key point in y, y_i' is the i-th key point in y', N is the number of landmark points, and d_{ipn} is the distance between the pupils of the eyes.

3) PEAK SINGLE-TO-NOISE RATIO

The peak single-to-noise ratio (PSNR) is used to calculate the ratio of the maximum possible power of a signal to the power of destructive noise that affects the accuracy of its representation. The equation of PSNR is defined as follows:

$$PSNR = 20 \times \log_{10}\left(\frac{255}{MSE}\right), \tag{11}$$

where y is the ground-truth image, y' is the generated image, y_i is the i-th pixel of the ground-truth image, y_i' is the i-th pixel of the generated image, N is the number of pixels in the image, and

$$MSE = \sum_{i=1}^{N} \frac{\left(y_i - y_i'\right)^2}{N}. \tag{12}$$

4) STRUCTURAL SIMILARITY

Structural similarity (SSIM) is used to measure the similarity between two images, the equation of SSIM is defined as:

$$SSIM\left(x, y\right) = \left[l\left(x, y\right)\right]^{\alpha}\left[c\left(x, y\right)\right]^{\beta}\left[s\left(x, y\right)\right]^{\gamma}, \tag{13}$$

where α, β, and γ are the weights used to adjust the three comparisons and are usually set to 1, and

$$l\left(x, y\right) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1}, \tag{14}$$

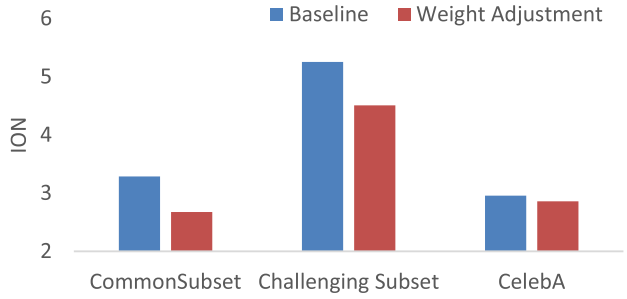

FIGURE 14. IONs comparisons on different datasets for landmark prediction.

TABLE 3. Improvements of IONs and IPNs on different experimental data.

Measure	Common Subset	Challenging Subset	CelebA
ION	19%	14%	3%
IPN	17%	14%	4%

$$c(x, y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2}, \tag{15}$$

$$s(x, y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3}. \tag{16}$$

In Equations (14)-(16), c_1, c_2 and c_3 represent constants, respectively used to avoid system errors caused by zero denominators. Moreover, μ_x and μ_y represent the means of x and y, respectively, σ_x and σ_y represent the standard deviations of x and y, respectively,

5) PARAMETER SETTINGS FOR FACE LANDMARK PREDICTION

To approximate the optimal setting of α in Equation (3), we conducted experiments to analyze the effectiveness of the loss function with new weight λ for training a face-landmark prediction model. We trained the models using EfficientV2-S as the main backbone and set the learning rate as 0.005, batch size as 8, and epochs as 25. Figures 12-13 show the experimental results for different settings, datasets and measures. Note that, the 300W is further decomposed into 2 subsets, namely Common Subset and Challenging Subset. From these results, we can know that almost the best settings of α are 4 whatever the dataset is.

C. EVALUATIONS FOR FACE-LANDMARK PREDICTION

1) IMPACT OF WEIGHT ADJUSTMENT FOR FACE LANDMARK PREDICTION

Based on the evaluation results for parameter α, the next result to investigate is the impact of weight adjustment. Figures 14 and 15 show the comparisons on different datasets for landmark prediction in terms of ION and IPN, respectively. In overall, the proposed method performs better than the baseline where the improvements are shown in Table 3.

Note that, the baseline indicates the EfficientV2-S without adjusting the weight in loss function for face landmark prediction. On the contrary, the called weight adjustment indicates that with a weighted loss function.

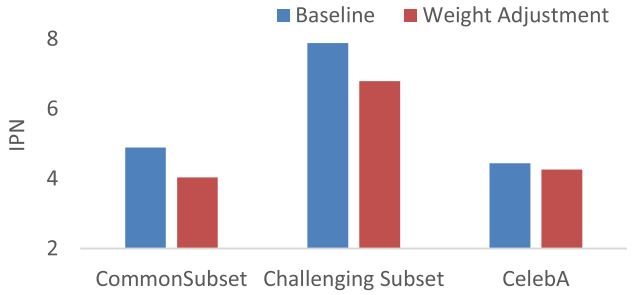

FIGURE 15. IPNs comparisons on different datasets for landmark prediction.

2) IMPACT OF VIEW DEGREE FOR FACE LANDMARK PREDICTION

The other evaluation to show is the impact of view degree for face landmark prediction. Table 4 shows the comparisons of IONs on different experimental datasets based on view degrees., which delivers some aspects. First, the proposed method outperforms the baseline whatever the data and view-degrees are. Second, the larger the view-degree, the larger the improvements. Third, the best performance occurs in the range of 0-40. In overall, even the face direction is not frontal, the proposed method still achieves obvious improvements.

D. EVALUATIONS FOR FACE INPAINTING

In the experiment of face inpainting, we first predict the face landmarks of the masked images using the weighted loss function with $\alpha = 4$ and the augmentation module. Next, the predicted face landmarks are concatenated with the occlusion images to form the conditions. Then, the landmarks are input into the face-inpainting model to complete the face. In these operations, we also perform the symmetry processing on the face and finally post-process the completed face to reduce the noises generated by complementation. We used SSIM to evaluate the proposed method in comparison with the competitor LaFin [38] for face inpainting on the dataset CelebA.

Figures 16 and 17 show the comparisons between the proposed method and compared method LaFin in terms of PSNR and SSIM, respectively on the dataset CelebA for face inpainting. Clearly, the proposed method is superior to the competitor. To realize the detailed results, the further analysis is shown in Figures 18 and 19. From these results, we can know that the proposed methods achieve better effectiveness than the compared method for mask-ablation experiments, regarding eyes, nose and mouth.

E. DEMONSTRATIVE EXAMPLES OF EXPERIMENTAL RESULTS

In order to visualize the real results, Figures 20 and 21 show the demonstrative examples of experimental results in terms of face-landmark prediction and face inpainting, respectively. From Figure 20, we can observe that the prediction results of

TABLE 4. Comparisons of IONs on different experimental datasets based on view degrees.

Range of View Degrees θ	Comparative Setting	Common Subset	Challenging Subset	CelebA
0-20	Baseline	3.241	4.901	2.379
	Weight Adjustment	2.661	4.484	2.349*
	Improvement	*18%*	*9%*	*1%*
21-40	Baseline	3.297	5.142	3.178
	Weight Adjustment	2.65*	4.35*	3.116
	Improvement	*20%*	*15%*	*2%*
41-60	Baseline	3.507	5.538	4.843
	Weight Adjustment	2.837	4.653	4.482
	Improvement	*19%*	*16%*	*7%*
61-90	Baseline	3.405	5.586	8.573
	Weight Adjustment	2.665	4.645	7.955
	Improvement	*22%*	*17%*	*7%*

Note: * indicates the best performances for each dataset.

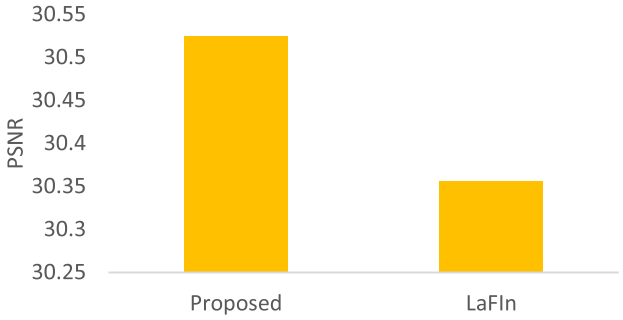

FIGURE 16. PSNR comparisons on the dataset CelebA for face inpainting.

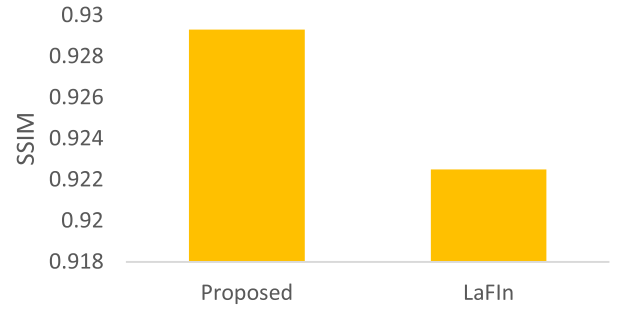

FIGURE 17. SSIM comparisons on the dataset CelebA for face inpainting.

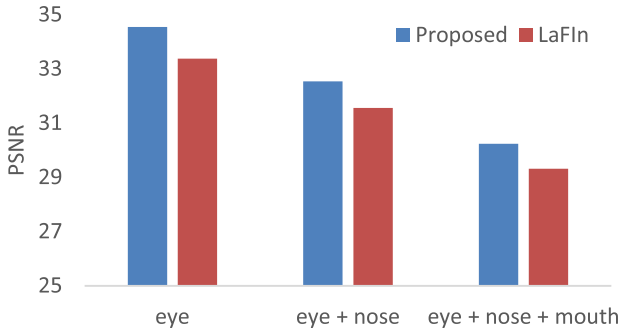

FIGURE 18. PSNR comparisons on the dataset CelebA under the different masks.

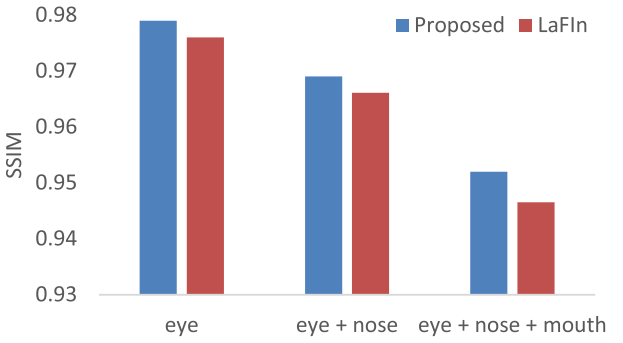

FIGURE 19. SSIM comparisons on the dataset CelebA under the different masks.

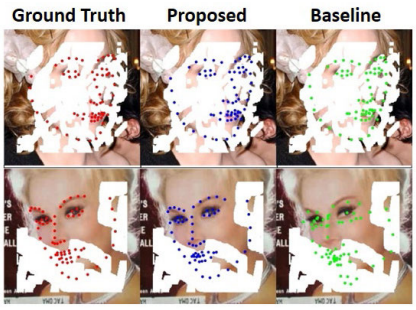

FIGURE 20. Demonstrative examples of experimental results for face landmark prediction.

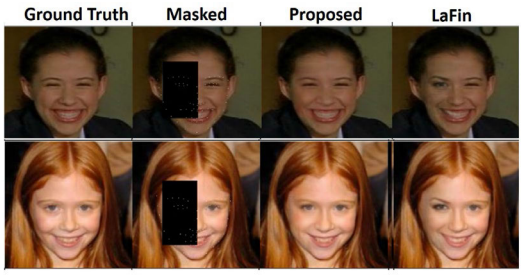

FIGURE 21. Demonstrative examples of experimental results for face inpainting.

the model are roughly close to the ground truth for images with small occlusion areas, while larger prediction errors usually appear in the face contour. Although there are often mistakes in the points on the face contour, it will not have much impact on the subsequent face inpainting task because

the face contour presented by connecting the predicted landmark is also very close to the original face. Furthermore, the proposed method performs better than the compared method. Even for images with large occlusion areas, as long as the facial features are not completely occluded, our proposed method can still perform well and preserve the geometry of the face. Figure 21 shows that the proposed method brings out more promising results than the compared method. It delivers an aspect that, applying our model for face restoration can maintain the symmetry of facial features. Even the occlusion size accounts for less than 10% of the total image area, the completed image is still very similar to the original image. The restored facial features are also very similar to the preserved ones.

V. CONCLUSION AND FUTURE WORKS

Image inpainting is also known as image completion, image hole filling and so on. The goal of image inpainting is to fill missing areas with natural and plausible contents. Face inpainting is one of image inpainting and the related goal is the same as that of image inpainting. In general, the application of face inpainting includes identifications of photos in ID card and images in surveillance videos. For example, if the image face in surveillance videos is not clear, you can recover the image by face inpainting. Therefore, the goal of this paper is to inpaint the polluted face photo by face landmarks and conditional generative adversarial networks. To reach this goal, in this paper, we propose a two-stage approach to achieve the face inpainting, including face landmark prediction stage and face inpainting stage. The main intents are to improve the symmetry of resulting images and to enhance the robustness of the prediction model. In the first stage, the masked image and the landmark with view degrees are used to train and predict landmarks for training and testing, respectively. In the second stage, the landmarks are used as conditions to generate the synthetic image by CGAN. The experimental results reveal that, the proposed method can better maintain the geometric structures and symmetric outlooks of inpainted faces than the compared method.

In the future, there remain a number of works to do. First, we will focus on these specific cases for face completion. Depending on the view degree of the face, the area of the left and right half of the face is also different. However, the facial features belonging to the same person still need to have the same attributes, such as facial makeup and the direction of the line of sight. Therefore, in the future, we will study how to train a model that can extract features such as facial makeup and sight direction and then use the method of contrastive learning to enhance the distinction of facial features. Second, the proposed method will be extended to improve the restoration ability of the image with a higher view degree. For face-landmark prediction, the environmental factors of the image can be considered more, and an adaptive noise reduction method can be added to speed up the convergence of the model and improve its robustness. Third, the landmark plays an important role for inpainting in this paper. From the experimental results, we can know that the inpainting performs effectively. This indicates the errors are very low, which can be reached in Figures 12-13. However, there are few unsatisfactory results still. These unsatisfactory instances are caused by that, the missing areas are so big that the landmark cannot be located perfectly. In the future, we will aim at this issue by looking for the better solutions.

ACKNOWLEDGMENT

This work was supported by the National Science and Technology Council, Taiwan, R.O.C., under Grant NSTC 113-2622-E-390-002 and Grant NSTC 112-2622-E-390-004.

REFERENCES

[1] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "PatchMatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, no. 24, pp. 1–11, Jul. 2009.

[2] J. Cai, H. Han, S. Shan, and X. Chen, "FCSR-GAN: Joint face completion and super-resolution via multi-task learning," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 2, no. 2, pp. 109–121, Apr. 2020.

[3] L. Cai, H. Gao, and S. Ji, "Multi-stage variational auto-encoders for coarse-to-fine image generation," in *Proc. SIAM Int. Conf. Data Mining*, Calgary, AB, Canada, May 2019, pp. 630–638.

[4] A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1200–1212, Sep. 2004.

[5] R. Dey and V. N. Boddeti, "3DFaceFill: An analysis-by-synthesis approach to face completion," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa, HI, USA, Jan. 2022, pp. 1224–1233.

[6] Q.-V. Dang and G.-S. Lee, "Scene text segmentation via multi-task cascade transformer with paired data synthesis," *IEEE Access*, vol. 11, pp. 67791–67805, 2023.

[7] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *Proc. 28th Annu. Conf. Comput. Graph. Interact. Techn.*, Los Angeles, CA, USA, Aug. 2001, pp. 341–346.

[8] A. A. Efros and T. K. Leung, "Texture synthesis by non-parametric sampling," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, Kerkyra, Greece, 1999, pp. 1033–1038.

[9] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2414–2423.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. 27th Int. Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2014, pp. 2672–2680.

[11] T.-P. Hong, J.-H. Wu, J.-H. Su, and T.-K. Yin, "Effective face inpainting by conditional generative adversarial network," in *Proc. IEEE Int. Conf. Big Data (BigData)*, Sorrento, Italy, Dec. 2023, pp. 6046–6050.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[13] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 5967–5976.

[14] J. Johnson, A. Alahi, and F. F. Li, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 694–711.

[15] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 12, pp. 4217–4228, Dec. 2021.

[16] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[17] Y. Li, S. Liu, J. Yang, and M. H. Yang, "Generative face completion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 5892–5900.

[18] G. Liu, F. A. Reda, K. J. Shih, T. C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 2018, pp. 1–16.

[19] H. Liu, B. Jiang, Y. Xiao, and C. Yang, "Coherent semantic attention for image inpainting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4169–4178.
[20] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Washington, DC, USA, Dec. 2015, pp. 3730–3738.
[21] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2813–2821.
[22] M. Mathieu, C. Couprie, and Y. Lecun, "Deep multi-scale video prediction beyond mean square error," in *Proc. Int. Conf. Learn. Represent.*, San Juan, Puerto Rico, May 2016, pp. 1–13.
[23] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.
[24] H. Nguyen-Truong, K. N. A. Nguyen, and S. Cao, "SRGAN with total variation loss in face super-resolution," in *Proc. 7th NAFOSTED Conf. Inf. Comput. Sci. (NICS)*, Ho Chi Minh City, Vietnam, Nov. 2020, pp. 292–297.
[25] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.
[26] G. Perarnau, J. van de Weijer, B. Raducanu, and J. M. Álvarez, "Invertible conditional GANs for image editing," 2016, *arXiv:1611.06355*.
[27] S. Qian, K.-Y. Lin, W. Wu, Y. Liu, Q. Wang, F. Shen, C. Qian, and R. He, "Make a face: Towards arbitrary high fidelity face manipulation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10032–10041.
[28] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: A styleGAN encoder for image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 2287–2296.
[29] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning Internal Representations by Error Propagation*. Cambridge, MA, USA: MIT Press, 1987, pp. 318–362.
[30] S. Rifai, P. Vincent, X. Müller, X. Glorot, and Y. Bengio, "Contractive auto-encoders: Explicit invariance during feature extraction," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2011, pp. 833–840.
[31] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 397–403.
[32] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "A semi-automatic methodology for facial landmark annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Portland, OR, USA, Jun. 2013, pp. 896–903.
[33] M. Tan and Q. Le, "EfficientNetV2: Smaller models and faster training," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2021, pp. 10096–10106.
[34] C.-T. Tu and Y.-F. Chen, "Facial image inpainting with variational autoencoder," in *Proc. 2nd Int. Conf. Intell. Robotic Control Eng. (IRCE)*, Singapore, Aug. 2019, pp. 119–122.
[35] A. Van-Den-Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 6309–6318.
[36] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn. (ICML)*, New York, NY, USA, 2008, pp. 1096–1103.
[37] M. Wang, "A novel image inpainting method based on image decomposition," *Proc. Eng.*, vol. 15, pp. 3733–3738, Jan. 2011.
[38] Y. Yang, X. Guo, J. Ma, L. Ma, and H. Ling, "LaFIn: Generative landmark guided face inpainting," 2019, *arXiv:1911.11394*.
[39] Y. Zeng, J. Fu, H. Chao, and B. Guo, "Learning pyramid-context encoder network for high-quality image inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 1486–1494.
[40] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 5908–5916.
[41] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2242–2251.

TZUNG-PEI HONG (Senior Member, IEEE) received the B.S. degree in chemical engineering from National Taiwan University, in 1985, and the Ph.D. degree in computer science and information engineering from National Chiao-Tung University, in 1992. He was with the Department of Computer Science, Chung-Hua Polytechnic Institute, from 1992 to 1994, and the Department of Information Management, I-Shou University, from 1994 to 2001. He was in charge of the whole computerization and library planning and preparation with the National University of Kaohsiung, from 1997 to 2000. He was the first Director of the Library and Computer Center, National University of Kaohsiung, from 2000 to 2001; the Dean of Academic Affairs, from 2003 to 2006; the Administrative Vice President, from 2007 to 2008; and the Academic Vice President, in 2010. He is currently a Distinguished Chair Professor with the Department of Computer Science and Information Engineering and the Department of Electrical Engineering and the Director of the AI Research Center, National University of Kaohsiung, Taiwan. He is also a joint Professor with the Department of Computer Science and Engineering, National Sun Yat-sen University, Taiwan. He received the First National Flexible Wage Award from the Ministry of Education in Taiwan.

JIN-HANG WU received the B.S. degree from the Department of Computer Science and Information Engineering, Tatung University, in 2020, and the B.S. degree from the Department of Computer Science and Information Engineering, National University of Kaohsiung, in 2022.

His research interests include data mining and multimedia information retrieval.

JA-HWUNG SU received the Ph.D. degree from the Department of Computer Science and Information Engineering, National Cheng Kung University, in 2010. Currently, he is an Assistant Professor with the Department of Computer Science and Information Engineering and the Chief Executive Officer of the AI Research Center, National University of Kaohsiung, Taiwan. He has held more than 17 patents in the USA and China and published more than 80 research papers in some premier journals and international conferences. Also, he served on the program committees and the reviewers in these journals and international conferences. His research interests include machine learning and multimedia information retrieval.

TANG-KAI YIN was born in Tainan, Taiwan, in 1967. He received the B.S. degree in electrical engineering from National Taiwan University, in 1990, and the M.S. degree in electrical engineering and the Ph.D. degree in electrical and computer engineering from Purdue University, West Lafayette, IN, USA, in 1994 and 1996, respectively. In 1999, he joined the Department of Information Management, Chia Nan University of Pharmacy and Science, Tainan, as an Assistant Professor. Since 2004, he has been with the Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung, Taiwan, where he was an Associate Professor, in 2006, and the Department Chair, from 2015 to 2018. His current research interests include computer vision, image processing, and deep learning both in advanced driver assistance systems and medical image processing.