E2F-GAN: Eyes-to-Face Inpainting via Edge-Aware Coarse-to-Fine GANs

Ahmad Hassanpour1*, Amir Etefaghi Daryani2*, Mahdieh Mirmahdi3*, Kiran Raja1, Bian Yang1, Christoph Busch1, and Julian Fierrez4

1Department of Information Security and Communication Technology NTNU, Gjovik, Norway
2Department of Electrical Engineering Amir Kabir University of Technology, Tehran, Iran
3Department of Artificial Intelligence and Robotics, University of Isfahan, Isfahan, Iran
4School of Engineering, Universidad Autonoma de Madrid, Madrid, Spain

Corresponding authors: Ahmad Hassanpour (e-mail: Ahmad.hassanpour@ntnu.no)

This work has been supported by project PRIMA (H2020-MSCA-ITN-2019-860315). J.F is also supported by project BBforTAI (PID2021-127641OB-I00 MICINN/FEDER).

ABSTRACT

Face inpainting is a challenging task aiming to fill the damaged or masked regions in face images with plausibly synthesized contents. Based on the given information, the reconstructed regions should look realistic and more importantly preserve the demographic and biometric properties of the individual. The aim of this paper is to reconstruct the face based on the periocular region (eyes-to-face). To do this, we proposed a novel GAN-based deep learning model called Eyes-to-Face GAN (E2F-GAN) which includes two main modules: a coarse module and a refinement module. The coarse module along with an edge predictor module attempts to extract all required features from a periocular region and to generate a coarse output which will be refined by a refinement module. Additionally, a dataset of eyes-to-face synthesis has been generated based on the public face dataset called CelebA-HQ for training and testing. Thus, we perform both qualitative and quantitative evaluations on the generated dataset. Experimental results demonstrate that our method outperforms previous learning-based face inpainting methods and generates realistic and semantically plausible images. We also provide the implementation of the proposed approach to support reproducible research via (https://github.com/amiretefaghi/E2F-GAN).

INDEX TERMS

Face Inpainting; Generative Adversarial Networks; Image Inpainting

I. INTRODUCTION

Image inpainting is used to complete missing information or substituting undesired regions of pictures with conceivable and fine-grained content. It encompasses a wide extent of applications in fields of restoring harmed photos, editing pictures, removing objects, etc [1][2]. Many conventional methods typically use low-level and hand-crafted features from the corrupted input image and utilize the priors or additional data. By propagating the extracted features from visible and well-structured parts to the missing regions or by filling missed small areas by looking and melding comparative patches from the same or other images. In spite of the fact that these strategies have great effects in the completion of replicating structures, they are restricted by the accessible regions in an image and cannot create novel image substance. In recent years, learning-based strategies have been proposed to overcome these constrictions by utilizing huge volumes of training data [3][4]. Notably, despite of great achievements of learning-based methods in this task, they are limited by at least three challenges: the inpainted area should be (C1) semantically filled based on overall scene, (C2) continuously structured with unmasked regions, and (C3) visually realistic.

Recently, deep convolutional neural networks (CNNs) and generative adversarial networks (GANs), known as learning-based methods, have been widely used for various image inpainting tasks including removing objects, noises, texts, and masks. Based on convolutional neural networks (CNNs) and using encoder-decoder network structure several works have been proposed for image inpainting [5]-[8]. For instance, Sidorov and Hardeberg [6] proposed an encoder-decoder network for denoising, inpainting and super-resolution for noised, inpainted and low-resolution images. Zhu et al. [5] proposed a patch-based inpainting method for various deep learning (DL) modules that have been proposed recently. Coarse-to-fine based methods exploit one [9, 10, 11] or two [12, 13,14]-stage architecture to complete content formation and texture refinement. A one-stage architecture (also termed coarse-and-fine architecture) consists of two parallel branches, coarse and fine, that extract two kinds of information simultaneously, coarse and fine information. The missed
region, then, can be constructed from the extracted information. Alternatively, a two-stage architecture generates an intermediate coarse image after recovering structures in the first stage, and then feeds it to the second stage for improving the texture. Additionally, another category called structural guidance-based methods uses an assistance algorithm to provide more information for the main inpainting method. An edge and a contour generator have been used within a two-stage architecture in [15] and [16] respectively.

Although, it is worthy to mention that in face inpainting, besides the above-mentioned challenges (i.e., C1-C3), we are facing further requirements. Notably, a facial representation can be considered for the purpose of biometric recognition due to the special topology of different facial elements (i.e., forehead, eyes, eyebrows, nose, mouth, jaw, chin, cheek) and their distinctive characteristics [42]. Thus, revealing the hidden parts of a face by using other elements such that the topological face elements along with consistency in face attributes (e.g., demographic and other biometric information [43]) are preserved is a challenging task, yet it will have a strong impact on the feasibility of biometric recognition conducted by human experts (i.e. in forensic investigation [44]) or by machine learning [45] or hand-crafted algorithms [46]. Therefore, the requirements of face inpainting are as follows: R1) the face topological structure should be reconstructed so that all elements are placed in the right position semantically and continuously. For this, first, the shape of the face (oval shape, square shape, round shape, etc.) should be predicted. Then all other elements should be placed proportionally within the predicted frame. Additionally, to look more realistic, the head pose should be naturally aligned and integrated with other elements. These requirements are the main challenges (i.e., C1-C3) of every inpainting method modified for face inpainting solutions. Since the aim of this paper is a special case of face inpainting where a large region of the face except eyes is hidden, besides R1, two other requirements which make the inpainting task more challenging should be considered. R2) Researchers have found that the area of skin around the eyes is useful to determine soft biometric information such as age or gender [17,37]. The proposed inpainting model should utilize the color, texture, and size of eyes and eyebrows to estimate this kind of demographic attributes and inpaint other face elements according to the estimated features. R3) The proposed solution should preserve the identity-related biometric properties present in the eyes regions [38,18] when generating the full face [39]. Noteworthy, this eye region is demonstrated to encode a large part of the identity information present in the face [44] enabling both person recognition and fake face detection [40].

Additionally, it is worthy to mention that the hidden portion of the image can directly affect the performance of proposed solutions, and clearly large masks make meeting the referred requirements (i.e., R1-R3) more difficult. Considering this issue, the aim of this paper is to complete the face based on the eyes region (periocular region), our used mask type will cover most parts of the face.

In this paper, a novel DL-based architecture has been proposed such that it complies with the referred requirements (i.e., R1-R3, see Fig. 1). Therefore, our contributions and novelties can be summarized as follows:

- In this work, an effective end-to-end solution for reconstructing the face based on just the eyes region has been proposed. This innovative GAN-based architecture called E2F-GAN benefits from the advantages of coarse-to-fine, coarse-and-fine, and structural guidance-based architectures. The code for our proposed method is available in GitHub1.
- By using various loss functions during the training process [41], not only the quality of inpainted regions but also demographic and biometric features have been preserved and measured by several quantitative and qualitative evaluation metrics.
- A new dataset of masked faces called E2Fdb has been generated and made publicly available (same GitHub indicated before).
- In terms of selecting the most informative guidance-based method, we experimentally show that edges provide more structural and contextual information compared to landmarks.

II. RELATED WORKS

In eyes-to-face inpainting, a face (a raw image indicated by $I_{H\times W\times N}$, hereafter) is corrupted by a binary image mask ($M_{H\times W\times N}$), where $H$, $W$, and $N$ show the height, width, and number of channels of the image respectively, and the corrupted image will be shown by $I_m$ ($I_m = I \odot M$, where $\odot$
is the element-wise production). The inpainting model $H$ takes $I_{M}$ and $M$ as input, and its output, reconstructed face, should fulfill the $R_1$, $R_2$, and $R_3$ ($\ell = I$). The proposed inpainting methods use different architectures and various types of masks. In this section, we review recent face inpainting methods based on DL architectures and widely used mask types.

### A. FACE INPAINTING METHODS

Apart from traditional methods which utilize low-level features extracted from the same image or a group of images, the learning-based strategy is the main focus of recent proposed methods due to using high-level features that enable them to inpaint the damaged regions semantically. In the following, we review several learning-based existing works that attempted to inpaint corrupted faces, similar to the aim of this paper.

The coarse-to-fine structure has been used in recent face inpainting tasks. Li et al. [19] proposed a generative-based coarse-to-fine structure that benefits from an attention layer to capture long dependency between features to generate more realistic images. Yu et al. [13] uses a coarse-to-fine structure to inpaint free-form masks. In the same context, Liu et al. [12] proposed a coarse-to-fine architecture with a novel attention layer. Chen et al. [19] proposed a coarse-and-fine structure including a coarse network for extracting global semantic information and a fine network to extract multi-level local features. Besides the coarse-to-fine based strategies, another category so-called structural guidance uses additional information to assist the main inpainting module. Nazari et al. [15] leverage an edge generator first to recover the edges, and the corrupted image is fed to the image inpainting network along with predicted edges. Chen and Liu [16] use a dual branch network including texture and edge branches to extract features and recover structures and textures of missed regions. Some works estimate facial landmarks to assist the main inpainting network [20][21]. In this paper, we will take the advantages of different architectures, i.e., coarse-to-fine, coarse-and-fine, and structural guidance.

The above-mentioned methods produce a unique result per each input. On the other hand, some approaches inpaint the corrupted regions differently per each execution for each specific input. Zheng et al. proposed a Variational Auto-Encoders (VAEs)-based [22] dual pipeline including a reconstructive path that uses the ground truth to learn the prior distribution of missing regions and a generative path for which the conditional prior is connected to the distribution obtained in the reconstructive path. An unsupervised conditional framework based on generative adversarial networks for varied image inpainting that can learn conditional completion distribution has been proposed by Zhao et al. [23]. A similar approach using GANs to restore low quality face images was recently proposed in [47]. It should be noted that, in E2F-GAN, we need a unique output for each input even after several executions to fulfill the requirements $R_2$ and $R_3$.

### B. MASK COVERAGE

The used masks in face inpainting scenarios can be classified into two categories called free and fixed-form masks. In widely used free-form masks [8, 10, 15, 21, 22, 24, 26], there are irregular shapes randomly placed on the images (Fig. 2a). Instead, in the fixed-form masks [13, 21, 24, 25], regular shapes cover part of the images which are located on the images randomly or purposefully (Fig. 2b) [24][25]. Since the aim of this paper is to complete the face based on eyes, our used mask type is in the latter category with a large-size mask ($\approx 75\%$ of the face).

### III. PROPOSED METHOD

The overall network architecture of our proposed method, which is based on a coarse-to-fine architecture and includes two main modules called coarse and refinement, is shown in Fig. 3. Different from others [2,3,7,13,19], both modules (i.e., coarse and refinement) are GAN-based networks, therefore, each of which includes a generator and a discriminator. The coarse module, which comprises a generator called coarse generator ($C$), has a dual encoder that follows the coarse-and-fine structure to capture global semantic features and extract multi-level features from the eyes region. Besides this module, a GAN-based refinement module which consists of a refinement generator ($F$) and a discriminator ($D_2$) has been utilized to improve the coarse outputs. Intuitively, the refinement network sees a more completed scene than the masked images, so its network can learn better feature...
representations than the coarse network. Therefore, our end-to-end method includes two GAN-based modules which are training to generate the final result. In the following subsections, each module is described in detail. Notably, facial landmarks [21] or edges [15] are usually the most widely-used structural guidance in image inpainting tasks. In our proposed E2F-GAN, where the used mask covers most parts of the face, predicting both landmarks and edges is a challenging problem. As a consequence, our proposed method utilizes both landmarks and edges during our experiments, in an effort to use the most effective structure (e.g., landmarks or edges). For facial landmarks, we used the landmark prediction method proposed in [27] and for predicting edges, we use the edge predictor proposed by Nazari et al. [15]. Both methods have been trained again on our generated dataset that contains specific eye masks. As we will see in the experiments, our quantitative and qualitative metrics will show that the edge structural guidance provides more effective information for our coarse generator. Therefore, in our final setup we use edges generated by an edge predictor ($E_e'$) as structural guidance for $C$.

A. Coarse Module
The proposed GAN-based coarse module is responsible for extracting the required features from the masked image and generating the first coarse result. To do this, we designed the module with three submodules including edge predictor ($E_e$), coarse generator ($C$), and discriminator ($D_1$). The following, we explain the role of each network, its architecture, and the used loss functions.

1) COARSE GENERATOR
The coarse generator has the main responsibility for meeting the three requirements (i.e., $R1$-$R3$). Not only the biometric and demographic feature should be extracted from the periorcular region, but also the initial coarse prediction should look realistic, and semantically and continuously structured. This is achieved using three networks: two encoders so-called fine encoder ($E_f$) and pose encoder ($E_p$), and a decoder. The encoder $E_f$ deals with the finest features of $I_m$ and $E_p$ deals with the predicted structure of faces obtained from $E_e'$. Therefore, first $I_m$ is fed to $E_e$ to predict edges of visible and hidden regions ($I_e'$) and then $I_e'$ is concatenated with $I_m$ to fed $E_p$. This assists to predict the pose of different elements of the face. Additionally, $I_m$ will be fed to $E_f$ with the aim of extracting identity attributes. Finally, the decoder will predict and inpaint the hidden regions based on the two feature maps received from $E_f$ and $E_p$. In the following, we describe each of these networks and their roles in our scheme.

Fine Encoder. The aim of using this encoder is mainly to extract demographic (e.g., age, gender) and biometric properties (e.g., identity, skin color) from $I_m$. Therefore, the skin color around the eyes, wrinkles, the size of eyes and eyebrows, the distance between two eyes, and other possible properties should be considered. On the other hand, it should be noted that due to the high coverage ratio of $I_m$, $E_f$ is fed
with a lot of unusable information (the black region). To prevent deteriorating the quality of output and filter out these pixels, the first seven blocks of $E_f$ are configured as with gated convolutions (GC) [14]. These blocks contain parallel convolution layers with different sorts of activation functions which assist to extract an appropriate feature map and eliminate extracted features from the masked region. Then, three interleaved gated residual blocks (IGRB) [19] have been placed after GC blocks to extract multi-level features.

**Pose Encoder.** For extracting coarse structure and global semantics features, and consequently preserving the quality as well as the structure of the predicted face, an encoder called pose encoder ($E_p$) has been placed in the Coarse Module ($C$). It has been fed by concatenation of $I_{edge}$ and $I_m$. Doing this, a receptive field for recognizing face structures will be available for $E_p$. However, the inputs $I_{edge}$ and $I_m$ are both sparse. To extract a meaningful feature map, similar to [29], we used three spatial pyramid dilation blocks (SPD) after six convolution layers. Notably, SPD blocks contain parallel convolution layers with various dilation rates to extract a large receptive field from the given input image.

**Decoder.** To inpaint the coarse output based on features extracted by $E_p$ and $E_f$, a decoder including seven layers (one attention layer and six upsampling convolution layers) has been used. In common encoder-decoder approaches, the decoder receives features directly from the encoder but in our proposed method, the decoder receives two types of features including low-level features extracted by large receptive fields that may lack detailed information (i.e., the output of $E_p$), and high-level detailed features with a small receptive field (i.e., the output of $E_f$). Thus, we use a CSAB as the first layer of the decoder to discriminate the more effective features from others by assigning more weights.

**Channel and Spatial Attention Block (CSAB).** According to the outputs of $E_p$ and $E_f$, the input to the attention block contains two types of features: a) large receptive field that may lack detailed information and b) output of $E_f$, i.e., high-level detailed features with small receptive fields. We adopt the concatenating operation to aggregate these two types of features. On the other hand, we may achieve redundant information about multi-level contextual information and this situation will not be efficient for our goals. Thereby, as shown in Fig.3, we adopt a specific attention block called channel and spatial attention block (CSAB) [19] to assign more weight to important features [48] and alleviate the interference of redundant features by channel and spatial attention. Hence, attention block composes of two main attentions which we will introduce. Convolution operation leads to local contextual information. Discriminative features representation is essential for inpainting. We leverage the attention mechanism to fulfill this desire. The channel attention emphasizes interdependent feature maps by exploiting the dependencies between channels. Meanwhile, the spatial attention encodes a wide range of contextual dependency within each channel, thereby improving the overall representation capability by gaining mutual for similar features.

**B. REFINEMENT MODULE**

The coarse module’s output ($\hat{I}_c$) consists of face coarse structure including placed face elements, stated face pose, specified color skin, etc., suffering from fine details. To add more details to the $\hat{I}_c$, we propose a GAN-based refinement module.

1) **REFINEMENT GENERATOR**

Inspired by the U-Net architecture [28] and the refinement network proposed by [29], we proposed a more effective architecture by replacing some DL blocks with SPD and self-attention (SA) blocks which receive the concatenation of $\hat{I}_c$ and $I_{edge}$ as its input. We have adopted SPD blocks with four dilation rates in the middle of our architecture to extract features with various receptive fields from input images and then used SA blocks between middle layers. SA benefits from the concept of self-similarity, which is useful for reclaiming the reconstructed pattern based on the remaining ground truth in a masked image. As mentioned before, the duty of this stage is that it should improve fine details of images, hence, we use reconstruction and perceptual losses to adjust the fine details.

**C. DISCRIMINATOR**

To inpaint and generate more realistic high-quality faces, both coarse and refinement modules have been designed based on GAN structures, thus, two discriminators have the responsibility of evaluating the output of $C$ and $F$. The coarse module’s discriminator ($D_1$) receives $\hat{I}_c$ and consequently the refinement module’s discriminator ($D_2$) has been fed by $\hat{I}_f$.

We have combined the concept of SN-GAN [30] and PatchGAN [31] for these discriminators to distinguish real or fake images. Besides this combination, we have used the hinge adversarial loss function for our discriminators. These combinations and loss functions help us to train our discriminators faster and more stable, distinguishing real or fake images efficiently.

**D. E2F-GAN END-TO-END TRAINING**

The E2F-GAN model is trained in a supervised and end-to-end manner. We have defined four groups of loss functions [41] for various parts of our proposed method to achieve considerable results. To train $C$, we have utilized four specific loss functions including reconstruction loss, perceptual loss,
style loss, and adversarial loss; and just reconstruction and perceptual losses have been used for training $F$. With the aim of having an end-to-end training process, we define the total loss $L$ which consists of four groups of component losses as below:

$$
L = \lambda_{rec}(L_{rec}^f + L_{rec}^i) + \lambda_{perc}(L_{perc}^f + L_{perc}^i) + \\
\lambda_{style}L_{style} + (\lambda_{adv}^g L_{adv}^g + \lambda_{adv}^d L_{adv}^d).
$$

(1)

In the style, the formulation of the used losses and the notion behind each loss is described. The reconstruction loss ($L_{rec}$) or per-pixel loss measures the pixel-wise difference between the synthesized image and the ground truth image. This loss is essential for maintaining texture information. It is calculated as the L1-norm between $\hat{I}_g$ and the corresponding ground truth $I_g$. $L_{rec}$ is defined as follows:

$$
L_{rec}^i = \frac{1}{H \times W \times N} \sum_{i=1}^{N} \sum_{j=1}^{W} |\hat{I}_g(i,j) - I_g(i,j)|
$$

(2)

where $z$ is replaced with $c$ or $f$ depending on the $L_{rec}$ is used for $C$ or $F$, respectively.

It is worth to mention that, an element-wise loss cannot consider high-level semantics. Accordingly, recent research [19, 21, 22] suggests using perceptual distances based on a pre-trained network, VGG19 which was trained on the ImageNet. The perceptual loss ($L_{perc}$) measures the difference between features extracted from the various layers of the VGG19 network for $\hat{I}$ and its corresponding ground truth.

$$
L_{perc}^z = \sum_{l=1}^{L} \frac{|\phi_z - \varphi_g|}{N_{l} \times H_{l} \times W_{l}}
$$

(3)

where $\phi_z$ and $\varphi_g$ are extracted features from $\hat{I}$ and $I_g$ respectively, and $z$ is replaced with $c$ or $f$ depending on the $L_{perc}$ is used for coarse or refinement, respectively. We extract features from $L$ layers of the pre-trained network. relu1_1, relu2_1, relu3_1, relu4_1, and relu5_1 of the VGG19 utilized to calculate $L_{perc}$ as well as $L_{style}$ described below.

In order to provide richer texture, we also employ style loss ($L_{style}$). In style loss, a Gram matrix calculates the correlation between channels in a feature map. The style loss then calculates on the features map produced by the pre-trained VGG19 network.

$$
L_{style} = \sum_{l=1}^{L} \frac{1}{N_{l} \times H_{l} \times W_{l}} \left\| G_l(\hat{I}_g) - G_l(I_g) \right\|_F
$$

(4)

where $G_l(\cdot) = \varphi_l(\cdot)^T \varphi_l(\cdot)$ stands for the Gram matrix corresponding to $\varphi(\cdot)$.

For generative adversarial learning, our discriminators are trained to distinguish between generated images and ground truth images. On the other hand, the generators strive to cheat the discriminators by hardening that classification. We employ hinge loss to train our model, $L_{adv}^d$ and $L_{adv}^g$ computed as follows:

$$
L_{adv}^g = -E[D_1(C(I_m, 1_{edge}))] - E[D_2(F(\hat{I}_g, 1_{edge}))]
$$

(5)

$$
L_{adv}^d = E[\text{Relu} (1 + D_1(C(I_m, 1_{edge}))) + E[\text{Relu} (1 + D_2(F(\hat{I}_g, 1_{edge})))]
$$

(6)

As mentioned before, we combine the used loss functions with appropriate weights as follows: $\lambda_{rec} = 1, \lambda_{perc} = 0.1, \lambda_{style} = 250, \lambda_{adv}^g = 0.1, \lambda_{adv}^d = 1$.

IV. EXPERIMENTS AND DISCUSSION

In this section, we evaluate the E2F-GAN performance on a new generated face dataset (E2Fdb) based on CelebA-HQ. We compared our results with three other methods called EdgeConnect (EC) [15], Pluralistic Image Completion (PIC) [22], LaFin [21]. To have fair comparison, the three methods have been trained using the E2Fdb. For quantitatively measuring the performance difference among the methods, we employ several statistical metrics. Moreover, to measure the amount of preservation of demographic and biometric features, we calculate False Non-Match Rate between original and inpainted faces. Using a competitive face biometric matcher [49] based on ArcFace [36].

A. DATASETS

We conduct all experiments on our generated dataset called E2Fdb (available on project’s GitHub page)

[^1]: https://github.com/amiretefaghi/E2F-GAN

2 extracted from the well-known CelebA-HQ dataset [32, 49]. To extract the periocular region from each face image, the images are reshaped to size 256 × 256 and then by utilizing a landmark detector [27], eyes are detected, similar to [50]. Doing this, $M$ and $I_m$ are produced for each image. Moreover, we removed mislabeling samples including those eyes covered by sunglasses or faces that have more than 45 degrees in one angle (roll, pitch, yaw) leading to hiding one of the eyes by
using WHENet [33] algorithms. Finally, the total number of samples is 24,554 among which 22,879 will be used for the training process and the rest, which is 1,685 images, for the test.

B. EVALUATION METRICS

We evaluate the image inpainting performance of the proposed model using quantitative and qualitative comparisons. For quantitative comparison, two types of metrics called statistical and identity metrics have been measured. In the following, we describe each category and its corresponding metrics briefly.

1) STATISTICAL METRICS

We use five statistical metrics: $\ell_1$ loss, Peak Signal to Noise Ratio (PSNR), Structural Similarity (SSIM) [34], Frenchet Inception Distance (FID) [35], and Total Variation (TV). Notably, the $\ell_1$ loss shows the model’s reconstruction ability for images. PSNR measures the visibility of errors between the

![Quality comparison among PIC, EC, LaFIn, and our proposed method.](image)
ground truth $I_Y$ and image inpainting $I$ to evaluate the image quality. SSIM aims at estimating the perceptual changes in the structural information, which shows human’s subjective feelings more accurately than PSNR. FID is a widely used metric in the image generation field to measure the visual quality. TV assists to measure the amount of noise in the image by calculating the sum of the absolute differences for neighboring pixels.

2) **IDENTITY METRICS**
To measure the amount of preservation of demographic and biometrics characteristics after completing inpainting process, we calculate the False Non-Match Rate (FNMR). FNMR is the rate at which a biometric algorithm miss-categorizes two captures from the same individual as being from different individuals. Here, we assumed that $I$ and $I$ are two faces for the same individual and using ArcFace [36], we calculate the corresponding embedding vectors for each face, and finally calculate the cosine similarity between each pair. Finally, the FNMR for different thresholds is shown.

C. **COMPARISON WITH EXISTING WORK**
By using the above-mentioned metrics and presenting some outputs, the results of our proposed method have been qualitatively and quantitatively compared against three state-of-the-art approaches, named PIC, EC, and LaFin. We trained the three methods over our generated dataset (i.e., E2Fdb) according to the best configurations of each method mentioned in the corresponding paper. In the following subsections, we present the results.

1) **QUANTITATIVE COMPARISONS**
The results of the statistical metrics calculated on the validation set of E2Fdb including 1,675 samples are reported in Table I. As can be seen from the numbers in Table I, E2F-GAN is superior over PIC, LaFin, and EC in most metrics, except for the $\ell_1$ loss for which LaFin works slightly better. Overall, our E2F-GAN outperforms the others by large margins in terms of FID, SSIM, PSNR, and TV metrics. More specifically, our large margins in FID and TV metrics demonstrate that our method can inpaint the masked image with much higher quality compared to other methods. Moreover, FNMR has been measured for E2F-GAN and other three compared methods as shown in Fig. 4. For different thresholds, E2F-GAN has lower false non-match rate which shows the ability of our algorithm extracting identity information from the periocular region and transferring it to the reconstructed face. Notably, since the PIC method generates different outputs for a specific input, we executed this method five times and the best results have been reported.

2) **QUALITATIVE COMPARISONS**
Fig. 5 shows some faces generated by our model, PIC, EC, and LaFin. Our model is able to generate high quality results and a large fraction of face structures including face shape, nose, mouth, forehead, etc. are appropriately placed with a plausible size. Moreover, to compare the quality of results in terms of gender and skin color, we present different faces in Figs. 5 and 6. As it can be observed, the quality of PIC and EC is really

| Method      | FID ↑ | SSIM ↑ | PSNR ↑ | TV ↓ | $\ell_1$ Loss ↓ |
|-------------|-------|--------|--------|------|-----------------|
| PIC         | 57.02 | 0.41   | 11.19  | 8.50 | 50.37           |
| LaFin       | 63.16 | 0.47   | 13.18  | 6.89 | 40.94           |
| EC          | 70.63 | 0.42   | 12.67  | 5.27 | 121.08          |
| E2F-GAN (ours) | **46.39** | **0.51** | **13.66** | **0.02** | **41.54** |
low compared to our and LaFiN results. Therefore, although like EC we used edge predictor in our scheme, there is a large margin between our outcomes. Additionally, with aim of further investigation of the models’ outputs regarding age and gender prediction based on the periocular region, we presented some challenging examples in Fig. 6. That figure shows three faces including two elders (a man and a woman) and a young woman. As seen in those examples, E2F-GAN can assess the age based on periocular region and reconstruct the face with a reasonable quality.

V. ABLATION STUDY

In this section, firstly we qualitatively and quantitatively analyze the effect of three main components of our proposed model including the edge predictor, the refinement module, and the attention block. Table II reports statistical metrics indicating the degree of effectiveness each of the three components in the performance of E2F-GAN. Specifically, the refinement network is the most conspicuous one which benefits the model by providing conformity and consistency among face components and skin texture around the eyes, such as wrinkles and skin color. The edge guidance contributes to ensuring that the structure of the face is well-preserved (see Fig. 7). Visually, the effectiveness of the attention block may not seem tangible. However, the quantitative results demonstrate the advantages of attention block. We also compared the effect of edge and landmark predictors. As shown in Table III, the edge guidance provides better values in most quantitative metrics specially for SSIM metric.

Finally, Fig. 8 shows a few challenging examples for preserving the gender of the person based on the periocular region. Our observations show that E2F-GAN can preserve the gender of subjects with a high accuracy.

VI. CONCLUSION

The aim of this paper is a particular case of face inpainting where we try to reconstruct the face based on just using the periocular region. To do this, we presented E2F-GAN, a GAN-based architecture that benefits from the advantages of coarse-to-fine, coarse-and-fine, and structural guidance-based architectures for face inpainting. It includes three main modules for extracting face’s edges (edge predictor), coarse prediction of face elements (coarse generator) and refining the coarse predicted image (refinement generator). We analyzed E2F-GAN and compared it with other well-known face inpainting methods to measure the efficiency and quality.

| Edge Predictor | Attention Block | Refinement module | FID ↓ | SSIM ↑ | PSNR ↑ | TV ↓ | L1 Loss ↓ |
|----------------|----------------|-------------------|-------|-------|-------|------|-----------|
| ✗              | ✓               | ✓                 | 64.89 | 0.34  | 12.18 | 12.91| 45.99     |
| ✓              | ✗               | ✓                 | 50.12 | 0.48  | 13.24 | 6.33 | 43.97     |
| ✓              | ✓               | ✗                 | 75.21 | 0.46  | 13.22 | 5.49 | 42.13     |
| ✓              | ✓               | ✓                 | 46.39 | 0.52  | 13.66 | 0.02 | 41.54     |

Figure 7. Illustrative comparison of the effect of various parts on final output.
performance. For doing this, we modified a widely used face inpainting dataset called CelebA-HQ such that the whole face except the periocular region is masked and used for E2F-GAN input, calling the resulting dataset E2Fdb. Our proposed inpainting algorithm E2F-GAN and the used dataset E2Fdb are both available in the project GitHub3. Several qualitative and quantitative metrics have been measured during our experiments to show the performance of E2F-GAN in terms of preserving identity and non-identity features of each face after inpainting. Experimental results show that our method outperforms previous learning-based face inpainting methods and E2F-GAN in terms of preserving identity and non-identity features.

Future work includes analyzing biometric quality aspects of the resulting faces using recent objective measures [51,52]; analyzing [49] and reducing [53] undesired biases in the face generation process; and combining multiple face generation approaches for better outputs [48].

REFERENCES

[1] A. Criminisi, P. Perez and K. Toyama, “Region filling and object removal by exemplar-based image inpainting,” IEEE Transactions on Image Processing, vol. 13, no. 9, pp.1200-1212, 2004.

[2] O. Elharrousi, N. Almaadeed, S. Al-Maadeed, and Y. Akhari, “Image inpainting: A review,” Neural Processing Letters, vol. 51, no. 2, pp.2007-2028, 2020.

[3] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell and A. A. Efros, “Context encoders: Feature learning by inpainting,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2536-2544, 2016.

[4] H. Yamauchi, J. Haber, and H.P. Seidel, “Image restoration using multiresolution texture synthesis and image inpainting,” in Proceedings Computer Graphics International, pp. 120-125, 2003.

[5] X. Zhu, Y. Qian, X. Zhao, B. Sun, and Y. Sun, “A deep learning approach to patch-based image inpainting forensics,” in Signal Processing: Image Communication, vol. 67, pp.90-99, 2018.

[6] O. Sidorov and J. Y. Hardeberg, “Deep hyperspectral prior: Single-image denoising, inpainting, super-resolution,” in Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp. 3844-3851, 2019.

[7] Y. Zeng, J. Fu, H. Chao, and B. Guo, “Learning pyramid-context encoder network for highquality image inpainting,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1486-1494, 2019.

[8] H. Liu, B. Jiang, Y. Xiao and C. Yang, “Coherent semantic attention for image inpainting,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4170-4179, 2019.

[9] Z. Guo, Z. Chen, T. Yu, J. Chen, and S. Liu, “Progressive image inpainting with full-resolution residual network,” in Proceedings of the 27th ACM International Conference on Multimedia, pp. 2496–2504, 2019.

[10] J. Li, N. Wang, L. Zhang, B. Du and D. Tao, “Recurrent feature reasoning for image inpainting,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.7760–7768, 2020.

[11] H. Zhang, Z. Hu, C. Luo, W. Zuo, and M. Wang, “Semantic image inpainting with progressive generative networks,” in Proceedings of the 26th ACM International Conference on Multimedia, pp.1939–1947, 2018.

[12] H. Liu, B. Jiang, Y. Xiao, and C. Yang, “Coherent semantic attention for image inpainting,” in Proceedings of the IEEE International Conference on Computer Vision, pp. 4170–4179, 2019.

[13] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and Thomas S Huang, “Generative image inpainting with contextual attention,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5505–5514, 2018.

[14] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, “Free-form image inpainting with gated convolution,” in Proceedings of the IEEE International Conference on Computer Vision, pp. 4471–4480, 2019.

[15] K. Nazeri, E. Ng, T. Joseph, F. Qureshi and M. Ebrahimi, “EdgeConnect: Structure Guided Image Inpainting using Edge Prediction,” in IEEE/CVF International Conference on Computer Vision Workshop (iCCCVW), pp. 3265–3274, 2019.

[16] M. Chen and Z. Liu, “EDBGAN: Image inpainting via an edge-aware dual branch generative adversarial network,” IEEE Signal Processing Letters, 28, pp.842-846, 2021.

[17] E. Bobrov, A. Georgievskaya, K. Kiselev, A. Sevastopolsky, A. Zhavoronkov, S. Gurov, K. Rudakov, M.D.P.B. Tobar, S. Jaspers, and S. Clemann, “PhotoAgeClock: deep learning algorithms for development of non-invasive visual biomarkers of aging,” Aging (Albany NY), vol. 10, no. 11, p.3249, 2018.

[18] C. Rathgeb, C. Busch. Handbook iris and Periocular Biometric Recognition, Institution of Engineering and Technology (IET), 2017.

[19] M. Chen, Z. Liu. Ye, and Y. Wang, “Attentional coarse-and-fine generative adversarial networks for image inpainting,” Neurocomputing, 405, pp.259-269, 2020.

[20] L. Song, J. Cao, L. Song, Y. Hu, and R. He, “Geometry-aware face completion and editing,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 1, pp. 2506-2513, 2019.

[21] Y. Yang, and X. Guo, “Generative Landmark Guided Face Inpainting,” in Pattern Recognition and Computer Vision, 2020.

---

3 https://github.com/amiretefaghi/E2F-GAN
[22] C. Zheng, T.J. Cham, and J. Cai, “Pluralistic image completion,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1438-1447.

[23] L. Zhao, Q. Mo, S. Lin, Z. Wang, Z. Zuo, H. Chen, W. Xing, and D. Lu, “UCTGAN: Diverse image inpainting based on unsupervised cross-space translation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5741-5750.

[24] N. Ud Din, K. Javed, S. Bae and J. Yi, “A novel GAN-based network for unmasking of masked face,” IEEE Access, vol. 8, pp.44276-44287, 2020.

[25] Y. Zeng, J. Fu, H. Chao and B. Guo, “Learning pyramid-context encoder network for high-quality image inpainting,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp.1486-1494.

[26] 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 Proceedings of the European Conference on Computer Vision, 2018, pp.85-100.

[27] A. Bulat, and G. Tzimiropoulos, “How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks),” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp.1021-1030.

[28] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in Medical Image Computing and Computer-Assisted Intervention, 2015, pp.234-241.

[29] C.T. Li, W.C. Siu, Z.S. Liu, L.W. Wang, and D.P.K Lun, “DeepGIN: Deep Generative Inpainting Network for Extreme Image Inpainting,” in Computer Vision – ECCV 2020 workshops, Glasgow, UK, 2020.

[30] T. Miyato, T. Kataoka, M. Koyama and Y. Yoshida, “Spectral Normalization for Generative Adversarial Networks,” in ICLR, Vancouver, Canada, 2018.

[31] K. Lata, M. Dave and K. N. Nishanth, “Image-to-Image Translation Using Generative Adversarial Network,” in 3rd International conference on Electronics, Communication and Aerospace Technology, 2019, pp.186-189.

[32] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of gans for improved quality, stability and variation,” in ICLR, Vancouver, Canada, 2018.

[33] Y. Zhou, and J. Gregson, “WHENet: Real-time Fine-Grained Estimation for Wide Range Head Pose,” in The 31st British Machine Vision Conference(BMVC), 2020.

[34] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” IEEE Transactions on Image Processing, vol. 13, no. 4, pp.600-612, 2004.

[35] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs trained by a two time-scale update rule converge to a local Nash equilibrium,” in NeurIPS, 2017, pp.6626–6637.

[36] J. Deng, J. Guo, N. Xue and S. Zafeiriou, “ArcFace: Additive angular margin loss for deep face recognition,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp.4690-4699.

[37] A. Fernando, K. H. Diaz, Silvia Ramis, F. J. Perales, and J. Bigun, “Facial Masks and Soft-Biometrics: Leveraging Face Recognition CNNs for Age and Gender Prediction on Mobile Ocular Images,” IET Biometrics, vol. 10 , pp. 562-580, 2021.

[38] A. Fernando, K. Raja, R. Raghavendra, C. Busch, J. Bigun, R. Vera-Rodriguez, and J. Fierrez. “Cross-sensor periocular biometrics for partial face recognition in a global pandemic: Comparative benchmark and novel multialgorithmic approach.” arXiv preprint arXiv:1902.08123, 2019.

[39] F. Alonso-Fernandez, R. A. Farrugia, J. Fierrez and J. Bigun, “Super-Resolution for Selfie Biometrics: Introduction and Application to Face and Iris,” in Selfie Biometrics, Springer, 2019, pp.105-128.

[40] R. Tolosana, S. Romero-Tapiador, R. Vera-Rodriguez, E. Gonzalez-Sosa and J. Fierrez, “DeepFakes Detection across Generations: Analysis of Facial Regions, Fusion, and Performance Evaluation,” Engineering Applications of Artificial Intelligence, vol. 110, pp. 104673, April 2022.