Received July 26, 2021, accepted September 27, 2021, date of publication October 4, 2021, date of current version October 12, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3117718

Automated Artifact Retouching in Morphed Images With Attention Maps

GUIDO BORGHI, ANNALISA FRANCO, GABRIELE GRAFFIETI, AND DAVIDE MALTONI, (Senior Member, IEEE)

Dipartimento di Informatica—Scienza e Ingegneria (DISI), Università di Bologna, 47521 Cesena, Italy

Corresponding author: Guido Borghi (guido.borghi@unibo.it)

This work was supported in part by the European Union's Horizon 2020 Research and Innovation Program under Agreement 883356, and in part technical support provided by the Istituto Poligrafico e Zecca dello Stato.

ABSTRACT Morphing attack is an important security threat for automatic face recognition systems. High-quality morphed images, *i.e.* images without significant visual artifacts such as ghosts, noise, and blurring, exhibit higher chances of success, being able to fool both human examiners and commercial face verification algorithms. Therefore, the availability of large sets of high-quality morphs is fundamental for training and testing robust morphing attack detection algorithms. However, producing a high-quality morphed image is an expensive and time-consuming task since manual post-processing is generally required to remove the typical artifacts generated by landmark-based morphing techniques. This work describes an approach based on the Conditional Generative Adversarial Network paradigm for automated morphing artifact retouching and the use of Attention Maps to guide the generation process and limit the retouch to specific areas. In order to work with high-resolution images, the framework is applied on different facial crops, which, once processed and retouched, are accurately blended to reconstruct the whole morphed face. Specifically, we focus on four different squared face regions, *i.e.* the right and left eyes, the nose, and the mouth, that are frequently affected by artifacts. Several qualitative and quantitative experimental evaluations have been conducted to confirm the effectiveness of the proposal in terms of, among the others, pixel-wise metrics, identity preservation, and human observer analysis. Results confirm the feasibility and the accuracy of the proposed framework.

INDEX TERMS Automated artifact retouching, conditional generative adversarial networks, deep neural networks, face morphing, single-image morphing attack detection.

I. INTRODUCTION

The results of public evaluation campaigns [1] confirm that *Face Recognition Systems* (FRSs) are able to achieve impressive levels of accuracy, especially when operating in controlled scenarios. Unfortunately, several recent studies also confirm that digital image manipulations can severely affect FRS performance: this is especially true for the so-called face morphing attack [2], where face images of two individuals, usually referred to as *criminal* and *accomplice*, are mixed to produce a new image (*morphed face*) containing facial features that belong to both subjects. In this way, the two subjects can share the same legal document, *e.g.* the ID card or the passport, and, in particular, the criminal can elude identity-based controls. Nowadays, face morphing is considered one of the major security threats [3], particularly in the context of electronic identity documents, such as the *electronic Machine Readable Travel Document* (eMRTD), where it can be successfully exploited for criminal intents, for instance, to fool *Automated Border Control* (ABC) systems thus overcoming security checks at the borders. Furthermore, the availability of a number of free or commercial software for face morphing generation makes the risk even more serious. For this reason, the research community is devoting significant efforts to the development of *Morphing Attack Detection* (MAD) algorithms [3], able to discriminate between bona fide (not manipulated) images and images generated by a morphing process.

Generally, the chances of success of a morphing attack depend on two key elements [4]:

- *Identity:* A morphed image should be successfully matched to both parent subjects;
- *Quality:* A morphed image should have a high quality, *i.e.* it should be free from any visible and non-visible

The associate editor coordinating the review of this manuscript and approving it for publication was Nikhil Padhi.

artifact typically generated by the morphing process that could be spotted by a human observer, for instance, a police officer, or a FRS.

Different morphing generation techniques can lead to different results in terms of image quality. Most of the existing algorithms are landmark-based: they perform a combination of image warping and texture blending, with respect to some reference points (*e.g.* eyes corner, nose, mouth, etc.) detected in the two images. As an alternative, some approaches based on *Generative Adversarial Networks* (GANs) [5] have been recently proposed [6], [7]. The advantage of landmark-based algorithms is that the degree of similarity with the two parent subjects can be easily controlled by modifying the morphing factor, which quantifies the presence of the two subjects in the morphing; on the other hand, the main limitation of these techniques is that visible artifacts are produced, in particular in the proximity of the main facial features (eyes, nose, mouth) due to insufficient or imprecise landmark positions detected. The generation of high-quality morphed images requires a tedious and time-consuming manual post-processing aimed at manually retouching the images to remove any visible artifact. On the contrary, GAN-based approaches usually overcome this limitation since the generated images are not affected by the presence of morphing artifacts, even though some specific GAN artifacts could arise [8]; their main limitation is that generating high-resolution images is quite complex, in terms of video memory requirements and training stability. In addition, the similarity with the parent subjects is more difficult to control and the resulting image is likely to fool automatic FRSs but not human experts [7].

This work represents a first investigation towards the definition of a morphing strategy able to combine the advantages of the two aforementioned morphing categories. Indeed, the underlying idea is to use a landmark-based approach to generate the morphed images and to delegate to a *Conditional GAN* (cGAN) [9] the subsequent post-processing stage aimed at removing the morphing artifacts. Then, the proposed method can exclude the need for human manual intervention on morphed images and can simplify the generation of large datasets of high-quality images to train and/or test MAD algorithms, especially if deep learning-based.

As visually summarized in Figure 1, the presented framework receives as input the morphed image and a related Attention Map, aimed at highlighting the image artifacts and computed through a logical operation on the warped images belonging to the accomplice and the criminal (see Section III-B). The morphed image is then cropped; each crop contains a portion of the face usually affected by artifacts: in this work, we focus on the left and right eyes, the nose, and the mouth areas. Then all the crops are improved by the cGAN and are blended in the initial morphed face, following the procedures analyzed in Section III-D, obtaining an image with a reduced amount of artifacts, in terms of ghosts, blurring, and texture inconsistency.

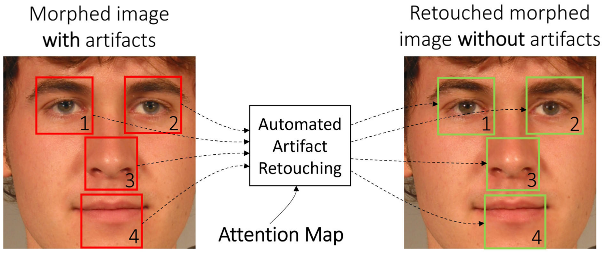

Morphed image with artifacts
Retouched morphed image without artifacts
1
2
3
4
Automated Artifact Retouching
Attention Map
1
2
3
4

FIGURE 1. **Simple outline of our task. Given as input the whole face, four different patches are cropped around the eyes, the nose and the mouth and the attention map is computed. The framework outputs retouched patches that are blended on the original input face.**

A detailed overview of the proposed framework and operations is reported in Figure 2.

To validate our approach, we conduct extensive analysis on the retouched morphed faces. Among the others, we exploit pixel-wise metrics in order to define the best approach to create the Attention Map and to quantify the overall quality also against competitors, as reported in Section IV-B. Then, in Section IV-C, we test the ability to preserve the identity of the retouched patches, and we show that the final retouched morphed faces are more realistic from the point of view of MAD algorithms and humans. The last part of the paper (Section V) draws some conclusions and highlights future research directions.

II. RELATED WORK

The treats to computer vision and biometric systems given by the face morphing attack was first described by Ferrara *et al.* [2], picturing a use case where a criminal wants to fool ABC gates at the airports. Afterward, face morphing has raised the attention of the research community [4], [10], since morphing attack represents a severe threat for all the applications that rely on automatic face verification algorithms. The chances of success for a morphing attack are also related to the image quality and the capability of automatically producing realistic morphed images is a desirable feature for any morphing pipeline since it will avoid the slow and tedious manual retouching.

Face automatic retouching was recently explored by Shafaei *et al.* [11] using GANs. Though the proposed method showed good results on a variety of facial imperfections, it was not developed to detect and correct the typical morphing artifacts, making a direct comparison with our proposal and [11] not trivial to pursue. Wang *et al.* [12] proposed a method to automatically detect image manipulations (mainly warping) inserted during the retouching of face photographs. The proposed approach is also able to undo the warp and reconstruct the original image without any modification. While the method produced good reconstructions, it cannot be used in our scenario, since the approach requires the optical flow of the image transformation, and we do not have access to such information.

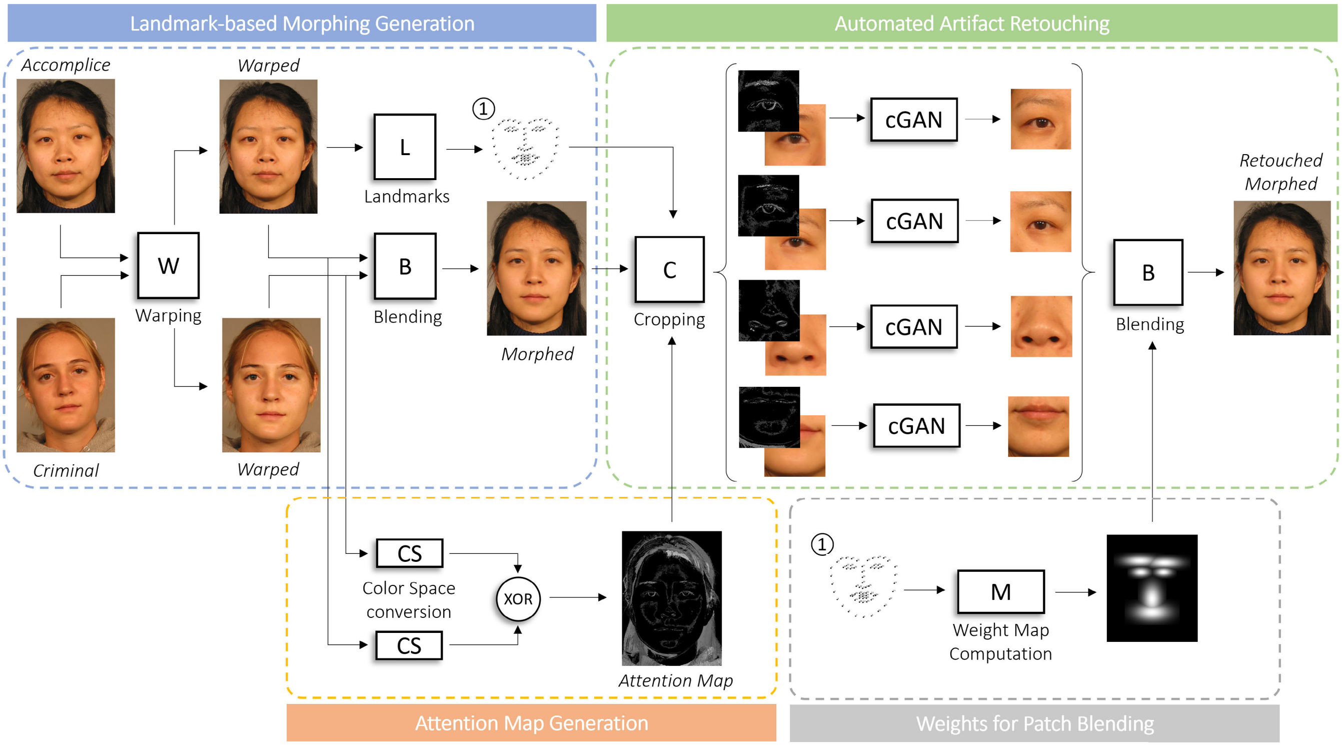


FIGURE 2. Overview of the proposed framework. In the blue rectangle, it is represented the morphing process that relies on the warping and blending operations. The two warped images are then used to compute the attention map (orange, Section III-B), while the landmarks detected on the accomplice warped face are used to crop the patches and to compute the weighted mask (grey, Section III-D) for the final blending. In the green rectangle, the core of the proposed system is shown: each patch is processed with the related attention map through a conditional GAN to produce a retouched patch (Section III-C), and all the generated patches are finally blended.

Apart from retouching, detecting if an image was forged is a complex problem per se. In recent years, many approaches have been proposed in the field of image forensic to recognize if an image was forged and to detect the forged regions. Bondi *et al.* [13] proposed a model to detect spliced images by looking at patches that come from different camera models. Another possible approach is to directly identify physical inconsistencies on different patches of the image [14]. Popescu *et al.* [15] discovered that common image forgery techniques may alter the underlining statistic of images, even without leaving any visual clue in the results. Their proposed method looks for these correlations in order to detect traces of image falsification. Another approach to identify forged images is based on the detection of inconsistent reflections and geometric inconsistencies [16]. However, the operations performed by morphing algorithms are often complex and could be difficult to model. Moreover, the discussed approaches only detect image manipulations but do not undo or correct them. Our approach, by contrast, is an end-to-end model that expressly detects and retouches the morphing artifacts.

One of the most similar works to our proposal was proposed by Seibold *et al.* [17], in which a style transfer approach, inspired by [18], is adopted for morph enhancement. The authors noted that the morphed images are often blurred and present fewer details than real face images (especially pores, scars, nevi, etc.). In this case, a pretrained CNN is used to extract style and content features from the contributing images and the morphing image respectively. Following the typical style transfer algorithm [18], the content will be preserved (so the morphed image should not change so much and the identities should be maintained) but the style of the resulting image will be more similar to a real image (sharper borders, enhanced texture, etc.). This procedure has the advantage of not requiring any training or fine-tuning on the particular dataset, making it a general strategy for face image enhancement. However, the objective, in this case, seems to be a more general improvement of the image texture, rather than an explicit removal of morphing artifacts. Moreover, the limited size of generated images and the impossibility to control the retouched areas represent the main limitations of this method, as described by the authors in the experimental evaluation of this paper.

Recently, GANs have been explored to create morphed images. In the work by Damer *et al.* [6] the two parent subjects are merged in the latent space, using an encoder-decoder architecture and a discriminator to enforce realism in the results. Though the results do not present artifacts due to the misalignment of landmarks, and the morphing can be done directly using the two images without any further pre-processing, the method in [6] produces images of dimension 64 × 64, which do not meet ICAO standards [19]. An improvement was recently proposed by Zhang *et al.* [7], using an architecture based on StarGAN [20]. As anticipated before, GAN-based methods do not produce the typical artifacts of landmark-based morphing algorithms, since those artifacts are the result of a misalignment or absence of facial landmark or different face geometries. However, GAN-based approaches also insert specific artifacts in the results [8]. Moreover, some parts of the face such as hairs or beards are usually blurred and flattened in a way that is easily recognizable by human observers. For these reasons, in our approach, we mainly focus on the correction and retouching of morph images produced by landmark-based algorithms.

Since morphed images represent a severe threat to security, a lot of research effort has been devoted to the detection of the Morphing Attack Detection (MAD) task. Generally, morphing attack detection can be divided into two branches [3], [21]: *Differential Morphing Attack Detection* (D-MAD) and *Single-image Morphing Attack Detection* (S-MAD). In the former method, the detection is based on two input images, and it consists of deciding if the first image (*probe*) is morphed or not, relying on the second image which is a securely live-captured image of the subject. In the latter methodology, only one image is available, so the morphing detection should be based solely on image analysis and statistics, assuming that the application of a morphing process leaves traces and anomalous patterns in images.

In this work, we focus on image retouching to improve the quality of morphed images, and the effectiveness of the proposed approach can be assessed also observing the effects on morphing attack detectors (MAD), for which the retouched images should represent a more difficult test. We concentrate on the S-MAD scenario since is more suitable for our data and operations, and it is more similar to how people, such as police officers, evaluate photographs during the issue of identity documents. S-MAD was addressed using a fusion of several classical image features, such as Histogram of Oriented Gradient (HOG) [22], Local Binary Patterns (LBP) [23], Binarized Statistical Image Features (BSIF) [24] and using a Support Vector Machine (SVM) [25] classifier with an RBF kernel [26]. A similar approach with different features and a fusion at the score level was proposed by Scherhag *et al.* [27]. An SVM was also used as a classifier by Zhang *et al.* [28], computing the Fourier spectrum of sensor pattern noise as image features. Raghavendra *et al.* [29] proposed to extract features using two different pretrained deep neural networks. The authors claimed that, even if the two networks were trained using the same dataset, the features are different and complementary. Following the feature extraction, the images are classified as morphed or real by a Probabilistic Collaborative Representation Classifier (P-CRC). In this work, we focus on two recent works. The first one is described in [30], in which an ensemble of features (HOG, LBP, BSIF) is extracted from different scales and color spaces of the same input images. The set of features is then classified through a Collaborative Representative Classifier (CRC) [31], obtaining state-of-art results in public evaluation benchmarks. The second one is reported in [32] and it is based on a deep learning-based paradigm, *i.e.* a deep neural network is trained and exploited in order to predict if a single input image is morphed or not. In Section IV-C we will use both methods to assess the quality of the morphed images retouched by the proposed framework.

III. PROPOSED METHOD

An overview of the proposed method is reported in Figure 2. As depicted, the goal of the whole framework is to output a retouched morphed image, *i.e.* an image without artifacts such as shadows, ghosts, double edges, blurring, and similar. The input of the core of the framework is represented by a morphed image and the related attention map. These two images are cropped relying on the landmark positions, they are divided into four patches (the right and left eyes, the nose, and the mouth) and used as input for each conditional GAN. The attention map generation procedure relies on the warped images of the accomplice and the criminal used during the morphing process and consists of a color space conversion followed by a bit-wise logical operation. Finally, the generated patches are merged through a weighted blending procedure into the original morphed face in order to obtain the final retouched morphed face.

Thus, from a formal point of view, the core of the proposed method consists in learning a function defined as:

$$\Psi : \mathbb{R}^{3\times w\times h} \oplus \mathbb{R}^{n\times w\times h} \rightarrow \mathbb{R}^{3\times w\times h} \tag{1}$$

where function Ψ takes as input an image with shape $(3+n)\times w\times h$, resulting from the concatenation along the channel dimension between an RGB image (3 channels) and an Attention Map (n channels), and outputs an intensity image in the RGB domain (3 channels). In this formulation, w and h represent the width and the height of both intensity images and Attention Maps.

A. PATCH PREPARATION

We empirically defined four main regions in which manual intervention is usually required to correct artifacts. These areas include the left and right eyes, the nose, and the mouth. The morphing process generally produces ghost artifacts also in the face surrounding region (i.e. hairs), but most morphing algorithms usually adopt automatic post-processing able to remove such defects by simply substituting the hair and background area with that of one of the parent images. For this reason, we will focus here only on the internal face region. The cropping procedure is based on landmarks extracted through the *DLib* library [33]. Therefore, the indexes of landmarks follow the convention proposed in [34]. Left and right eyes are cropped relying on a bounding box with top-left origin (x_B, y_B) and width w_B and height h_B computed as follows:

$$\begin{aligned} w_B, h_B &= (x_{27} - x_{23}) \times 1.2 \\ x_B &= \frac{x_{44} + x_{45}}{2} - \frac{w_B}{2} \\ y_B &= \frac{y_{44} + y_{45}}{2} - \frac{h_B}{2} \end{aligned} \tag{2}$$

The bounding box for the nose patch, instead, is defined as:

$$\begin{aligned} w_B, h_B &= (x_{36} - x_{32}) \times 1.6 \\ x_B &= x_{30} - \frac{w_B}{2} \\ y_B &= x_{30} - \frac{x_{31} - x_{29}}{2} \end{aligned} \tag{3}$$

Finally, in order to produce a square crop and to avoid the presence of facial details already included in the other patches, the mouth is cropped including a large portion of the chin:

$$\begin{aligned} w_B, h_B &= (x_{55} - x_{49}) \times 1.2 \\ x_B &= \frac{x_{51} + x_{53}}{2} - \frac{w_B}{2} \\ y_B &= \frac{x_{51} + x_{53}}{2} - \frac{x_{52} - x_{34}}{2} \end{aligned} \tag{4}$$

An example of resulting patches can be observed in Figure 1.

B. ATTENTION MAPS

The Attention Map is introduced to drive the retouch performed by the network and to focus its attention on the image regions affected by morphing artifacts. Conveying the modifications produced by the network to very small areas is fundamental to preserve the identity associated with the morphed image and to obtain high-quality results.

The map is built from data made available during the morphing algorithm described in [35], available from the Biometric System Laboratory website[1] and briefly recalled here. Given two images I_0 and I_1, the related facial landmarks P_0 and P_1, and the morphing factor $\alpha \in \mathbb{R}$, $0 < \alpha < 1$ representing the weight associated to image I_1, morphing can be defined as:

$$I_\alpha = (1-\alpha) \cdot w_{P_0 \to P_\alpha}(I_0) + \alpha \cdot w_{P_1 \to P_\alpha}(I_1) \tag{5}$$

Indeed, the morphing pipeline can be viewed as a combination of two operations:

- image warping, represented by the function $w_{P_i \to P_\alpha}$, needed to geometrically align the landmarks of the two input images to an intermediate position, according to the morphing factor α;
- texture blending, obtained as a weighted average of the pixels' intensity of the two warped images.

Given the two input images, properly aligned by warping, the artifacts are usually generated by the blending process, especially in those regions where the texture of the two images is noticeably different: this is common, for instance, in the pupil region or the eyelids, which are usually not perfectly aligned.

Under this assumption, the proposed Attention Map is derived by determining the texture details present in one of the two images and not in the other, by means of a pixel-level bit-wise logical exclusive-or operation (XOR, here referred with $\dot\vee$ symbol). The XOR operation is applied on a specific channel of the warped images converted to a given color space. Let's define $w_0 = w_{P_0 \to P_\alpha}(I_0)$ and $w_1 = w_{P_1 \to P_\alpha}(I_1)$ the two parent images properly warped according to the morphing factor α. The initial attention map is obtained as follows:

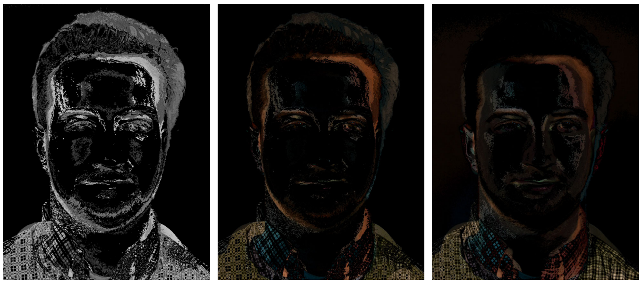

(a) M_1 (b) M_2 (c) M_3

FIGURE 3. Samples of attention maps computed using the procedures described in Section III-B. As depicted, each attention map focuses on different facial details and therefore leads to different final results, as reported in Table 2.

$$\mathcal{X}_c(w_0, w_1) = c(w_0) \mathbin{\dot\vee} c(w_1) \tag{6}$$

where c denotes the color space transformation and the selection of the channel of interest. The values too low ($<\tau_1$) or too high ($>\tau_2$) are discarded since they are usually related to negligible texture differences and lightning variations respectively. We will refer to the results of the thresholding operation as $\overline{\mathcal{X}}_c(w_0, w_1)$. In our experiments, we empirically fixed the two thresholds as follows: $\tau_1 = 50$, $\tau_2 = 200$.

With regard to the channel used for XOR computation, several alternatives have been evaluated and the most promising results have been observed for: i) RGB image converted to grayscale (average of the three channels); ii) X channel (related to red sensitivity) of the image converted to XYZ color space.

Also for the final attention map, two alternatives are proposed: i) grayscale map coinciding with $\overline{\mathcal{X}}_c$; ii) RGB map, obtained by a pixel-wise multiplication between $\overline{\mathcal{X}}_c$ and the difference between the two warped images $|w_0 - w_1|$, computed in the RGB color space and clipped to the proper range.

Three different combinations of color channels and gray/color output are evaluated in this work:

- Attention Map M_1:

$$M_1 = \overline{\mathcal{X}}_{RGB \to Gray}(w_0, w_1) \tag{7}$$

- Attention Map M_2:

$$M_2 = \overline{\mathcal{X}}_{RGB \to Gray}(w_0, w_1) \odot |w_0 - w_1| \tag{8}$$

- Attention Map M_3:

$$M_3 = \overline{\mathcal{X}}_{XYZ \to X}(w_0, w_1) \odot |w_0 - w_1| \tag{9}$$

[1] https://biolab.csr.unibo.it/research.asp?organize=Activities&select=&selObj=220

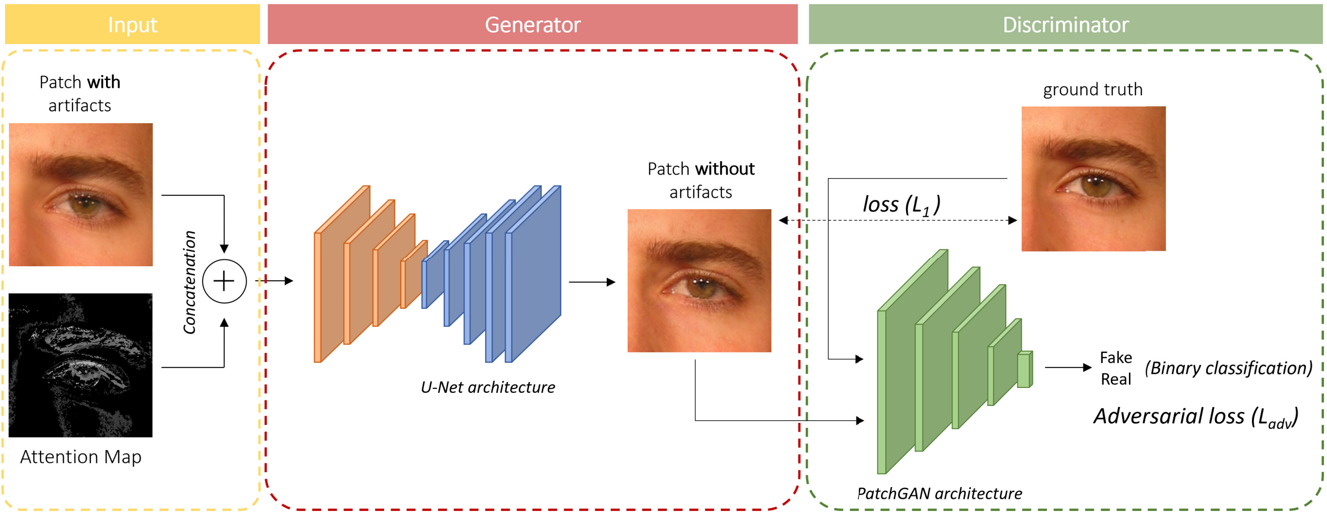


FIGURE 4. Overview of the conditional GAN paradigm exploited in the proposed framework. The input of the system is a concatenation (at channel level) between an RGB patch (here with the eye) and an attention map (see Section III-B). The generator network (in red) outputs the retouched image, while the discriminator network (in green) tries to classify whether the input image is generated (*fake*) or not (*real*).

In Equations 8 and 9, $|w_0 - w_1|$ denotes the absolute difference of the warped images. As shown in Figure 3, the three maps provide different information to the network; while M_1 only highlights the areas where morphing artifacts are more likely to appear (see Figure 3a), M_2 and M_3 provide additional color information that might be exploited by the network for artifact retouching (Figure 3b and 3c). A comparison between M_2 and M_3 can be interesting since M_2 is constructed by giving equal importance to all color components, while M_3 emphasizes the contribution of the red color, particularly relevant for some face parts (*e.g.* nose or mouth). The experimental results will provide a comparative evaluation of the three maps.

C. CONDITIONAL GAN

The core of the proposed framework is based on a Conditional GAN [36], *i.e.* a GAN in which the generation is conditioned through a given input image, as depicted in Figure 4. Taking inspiration from the work of Isola *et al.* [9], the architecture is based on a Generator network (here referred to as G) and a Discriminator network (referred to as D). In this way, G corresponds to the function Ψ (see Eq. 1) that estimates an image without visual artifacts, while D represents a function able to distinguish if an input image is “fake” (generated) or “real” (ground truth).

1) ARCHITECTURES

Network G is designed following the *U-Net* [37] architecture, *i.e.* a fully convolutional deep neural network with skip connections between the layer i and n_i, where n is the amount of total convolutional layers. The first part of the Generator acts as an encoder, mapping the input data in a 1024-dimensional embedding. Four convolutional layers with 128, 256, 512 and 1024 feature maps and kernel size of 5 (stride $s = 2$) are used. The *Leaky ReLU* [38] with a negative slope of 0.2 is used as activation functions and batch normalization is used to reduce the internal co-variance shift [39]. The latter part of the network is a decoder, able to generate images starting from the embedding space. The up-size procedure is based on four transposed convolutional layers with kernel size of 5 (stride $s = 2$) and 512, 256, 128 and 64 feature maps. In this case, the *ReLU* [40] is exploited as activation function.

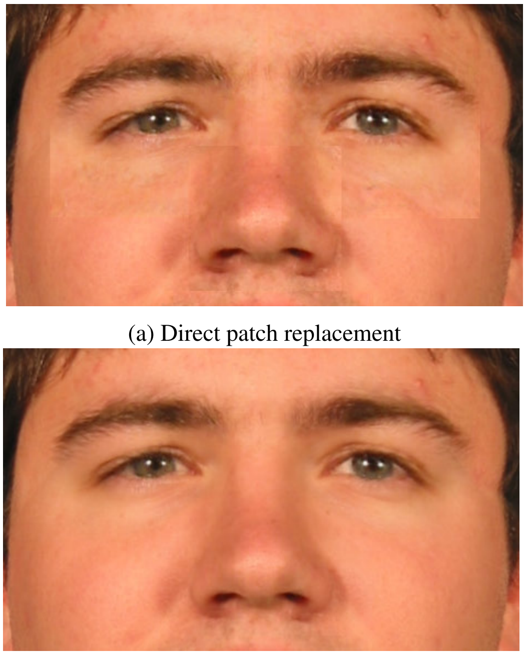
The discriminator network D predicts the probability of an input image to be a real or generated image. Following the principles presented in [9], [41], the D network is implemented as a convolutional *PatchGAN* classifier, which classifies 70 × 70 (in our experiments) patches belonging to the input image as generated or not. This classifier fits with our general goal since it is able to capture texture and style details [9], improving the overall quality of the generated images.

2) TRAINING

Networks G and D are trained following the well-known adversarial paradigm [5], based on the so-called min-max game. From a mathematical point of view, the training operation can be formalized as the optimization of the following problem:

$$\min_{\theta_g} \max_{\theta_d} \; \mathbb{E}_{x \sim p_{dpt}(x)}[\log(D(x))] + \mathbb{E}_{y \sim p_{gray}(y)}[\log(1 - D(G(y)))] \qquad (10)$$

in which $D(x)$ and $1 - D(G(y))$ are the probabilities of being a “real” or “fake” (generated) image, respectively. For the training of the discriminator network, we use the *Binary Cross Entropy* loss (L_{adv}), while for the generator we use a weighted combination of the adversarial loss, *i.e.* the opposite of the discriminator loss, and the L_1 loss to compute the final



(a) Direct patch replacement

(b) Weighted patch blending

FIGURE 5. Visual example of the weighted blending process. In (a) a direct patch replacement of the original image is applied (without weighted blending) and visible edges appear at the patch borders; in (b) the proposed weight map is applied for blending and the result is much smoother. (best in colors).

loss L_G as follows:

$$L_G = -L_{adv} + \lambda \frac{1}{N} \sum_{i=1}^{N} || G(I_i) - y_i ||_1 \tag{11}$$

where I and y are the input and the target image, respectively, and $\lambda = 100$. *Adam* [42] is used as optimizer for both the networks, with an initial learning rate $2 \cdot 10^{-4}$, $\beta_1 = 0.5$, $\beta_2 = 0.999$ and a batch size of 1. In all of our experiments, the size of the input images is first rescaled to 286×286 through a bicubic interpolation, and then randomly cropped to the final size of 256×256.

D. PATCH APPLICATION

Once the retouched patches have been generated, they have to be seamlessly integrated into the morphed image. This process is not trivial: a direct patch replacement, in fact, is unfeasible since some visible edges may arise at the borders as clearly shown in Figure 5a.

The solution adopted is an image blending that locally combines the original morphed image and the retouched patches on the basis of a weight map determined according to the local landmark density. Close to the landmarks, in fact, the contribution of the retouched patches to the final image must be predominant; as we move far away from such reference points, the weight associated with the original morphed image must increase. This blending operation guarantees a smooth transition and avoids the creation of visible edges (see Figure 5b).

The input for the retouched patch blending is the original morphed image I, and a set of generated patches $\mathcal{P} = \{P_i, i = 1 \ldots K\}$ where each patch $P_i = (b_i, L_i, R_i)$ is associated to the region of interest in the image b_i, the set of reference landmarks L_i and the retouched patch image R_i. The result of patch blending is the retouched morphed image $\tilde{I}$. The pseudocode of the algorithm is described in Algorithm 1; the reference points and distances used in the code are listed in Table 1, where the landmark used refer to the numbering convention proposed in [34].

Algorithm 1 Retouched Patches Blending

```
procedure ApplyPatches($I$, $\mathcal{P}$)
    $\tilde{I} \leftarrow I$
    for each $P_i \in \mathcal{P}, P_i = (\mathbf{b}_i, L_i, R_i)$ do
        $w \leftarrow width(\mathbf{b}_i), h \leftarrow height(\mathbf{b}_i)$
        Resize the patch image $R_i$ to size $w \times h$
        Initialize $M$ to an empty image of size $w \times h$
        for each $l_j \in L_i$ do
            $\mathbf{c} \leftarrow$ reference center $(c_x, c_y)$ for $l_j$ (see Table 1)
            $\mathbf{c}' \leftarrow$ map $\mathbf{c}$ to patch coordinates $\mathbf{b}_i$
            $d_x, d_y \leftarrow$ reference size for $l_j$ (see Table 1)
            $\sigma_x \leftarrow s \cdot d_x, \sigma_y \leftarrow s \cdot d_y$
            $G \leftarrow \mathcal{N}(\mathbf{c}', (\sigma_x, \sigma_y))$ of size $w \times h$
            $M \leftarrow$ element-wise maximum $(M, G)$
        end for
        Smooth borders in $M$
        $M \leftarrow M / \max(M)$
        $\tilde{I}[\mathbf{b}_i] \leftarrow M \cdot R_i + (1 - M) \cdot I[\mathbf{b}_i]$
    end for
end procedure
```

The size of the patches provided in output by the network is fixed (256×256) and an initial resize is needed to rescale the patch to the original size. Each patch $P_i \in \mathcal{P}$ is then applied to the unretouched morphed image by a local blending procedure based on a pixel-specific blending weight. The weight map is built as a multivariate Gaussian distribution reflecting the presence of landmark clusters in the patch (each patch may contain more than one landmark cluster, see Table1). For each cluster, in fact, the algorithm generates a specific Gaussian distribution $\mathcal{N}$ whose parameters (center and standard deviation) are determined on the basis of the position of the landmarks; for instance, for the nose patch, the highest weight is assigned to the nostrils area. A further Gaussian smoothing is finally applied to the whole patch weight map on the borders to further reduce the presence of the visible edges in the final image. After weight map computation, the weighted patch blending is finally executed (see Figure 2).

IV. EXPERIMENTAL RESULTS

In this section, we report several tests that have been carried out on the proposed framework. In Section IV-B, we analyze

TABLE 1. Reference landmarks for the different landmark clusters of the four patches. The given values are used to generate the Gaussian weight map associated to the landmark cluster: d_x and d_y represent, respectively, the reference width and height, while c_x and c_y are the coordinates of the central point (see Algorithm 1). Further details about landmarks indexes are reported in [34].

Patch	Cluster	d_x	d_y	c_x	c_y
Left Eye	Eyebrows	$x_{27} - x_{23}$	$\frac{(y_{23}-y_{25})+(y_{27}-y_{25})}{2}$	$\frac{x_{27}+x_{23}}{2}$	$\frac{y_{23}+y_{25}+y_{27}}{3}$
	Eye	$x_{46} - x_{43}$	$\frac{(y_{48}-y_{44})+(y_{47}-y_{45})}{2}$	$\frac{x_{46}+x_{43}}{2}$	$\frac{y_{44}+y_{45}+y_{47}+y_{48}}{4}$
Right Eye	Eyebrows	$x_{22} - x_{18}$	$\frac{(y_{18}-y_{20})+(y_{22}-y_{20})}{2}$	$\frac{x_{22}+x_{18}}{2}$	$\frac{y_{18}+y_{20}+y_{22}}{3}$
	Eye	$x_{40} - x_{37}$	$\frac{(y_{42}-y_{38})+(y_{41}-y_{39})}{2}$	$\frac{x_{40}+x_{37}}{2}$	$\frac{y_{38}+y_{39}+y_{41}+y_{42}}{4}$
Nose	Nose	$x_{36} - x_{32}$	$x_{34} - x_{30}$	x_{34}	y_{34}
Mouth	Mouth	$x_{55} - x_{49}$	$x_{58} - x_{52}$	$\frac{x_{55}+x_{49}}{2}$	$\frac{y_{51}+y_{53}+y_{57}+y_{59}}{4}$

and compare the results of the system in terms of pixel-wise metrics computed on single patches. This analysis allows to choose the best hyper-parameters and settings, such as the type of Attention Map, for the proposed system. Finally, in Section IV-C, we show how the framework performs on real morphed images, exploiting the weighted blending procedure to obtain a whole face starting from single retouched patches. This investigation is useful in order to assess the general quality of generated images, to provide quantitative results in terms of identity preservation and chances of deceiving morphing attack detection algorithms and human observers.

A. FRGC$_S$ AND FRGC$_M$ DATASETS

The training of deep neural networks and a quantitative evaluation of the proposed approach would require a large dataset of morphed images before and after manual post-processing. Unfortunately, such a dataset is not currently available. Therefore, we decided to generate images starting from the *Face Recognition Grand Challenge* (FRGC) dataset [44], chosen for the great variety of high-quality face images. For our experiments, we considered a subset of 1987 frontal images of different subjects, acquired in a controlled environment and with a uniform background. From this initial set of images, we create two datasets, here referred to as FRGC$_S$ and FRGC$_M$, respectively. For both datasets, the morphing algorithm used for image generation is described in Section III-B. In particular, we compute facial landmarks through the *DLib* libraries [33] and we use a morphing factor $\alpha = 0.5$.

FRGC$_S$ is intended for network training and quantitative evaluation of the results obtained, in order to define the experimental setting of the proposed framework in terms of selection of the best map type and the following ablation study. Since we have no pairs of morphed images before and after manual retouching (respectively with and without artifacts), we decided to simulate them. In particular, we generated the image with artifacts by morphing two images of the same subject, and we used as reference the first image used for morphing. The pairs thus obtained are reasonably similar to the real case since the difference between the two images is only related to morphing artifacts and not to a different identity (as it would happen if we morphed two different subjects). Some artifacts will be generated by the different subject's pose, however, to further increase the presence of artifacts, we applied a random perturbation to the landmarks of the first image used for morphing, thus explicitly causing a misalignment of the reference points. The perturbation applied consists of an affine transform including a random combination of rotation ($-5° \leq r \leq +5°$), translation ($-7\text{px} \leq t \leq +7\text{px}$) and scaling ($0.95 \leq s \leq 1.05$). This perturbation is the same for all the landmarks, to avoid the generation of unrealistic effects during morphing. Overall, we generated 4575 images, split into the training set (3555 image pairs of 196 subjects) and the test set (1020 image pairs of 32 subjects). Training and testing subjects (and datasets) are therefore disjoint. Image size in FRGC$_S$ varies from 864×648 to 1808×1356, with an inter-eye distance in the range $[168, 357]$. No compression is applied to the generated images.

FRGC$_M$ is designed to be used as a testing set for qualitative evaluation of the generated images as well as for a number of experiments related to identity preservation, human observer evaluation, and MAD testing. It has been generated by morphing images of two different individuals. In particular, for each of the 32 subjects (11 women, 21 men) in the testing set previously defined, four different images are morphed with one image of all the other subjects of the same gender (excluding symmetric pairs), thus leading to a total of 1060 morphed images. It is worth noting that, since each subject is mixed with all the other subjects of the same gender, the resulting dataset comprises morphed images of heterogeneous quality, thus making possible a more comprehensive analysis of the results. The images in FRGC$_M$ are uncompressed, with a size ranging from 992×744 to 1739×1304, and an inter-eye distance in the range $[195, 346]$.

B. FRAMEWORK ANALYSIS THROUGH FRGC$_S$

As mentioned before, FRGC$_S$ allows computing quantitative results due to the presence of a reference image. Therefore, we compare two images, *i.e.* the reference and the retouched patch, through a variety of pixel-wise metrics. We implemented the metrics described in literature works [43], [45], being aware that the evaluation of the visual quality of images is still an open problem, as highlighted in [46]. Specifically, we use the L_1 and L_2 distance, the absolute and square-root differences, the *Root Mean Square Error* (RMSE), and

TABLE 2. Pixel-wise metrics computed on generated images from $FRGC_S$ dataset to compare the performance of the framework using different attention map types in input. Further details about metrics are reported in Section IV-B and [43]. On the top, arrows indicate the positive changing direction, in which better performance corresponds to a variation that follows the arrow. At the bottom of the table, average values computed on all patch types are reported. As shown, M_2 represents the best choice.

Patch Type	Map Type	Norm ↓		Difference ↓		RMSE ↓			δ-metrics ↑			Indexes ↑		Perc. ↓
		L_1	L_2	Abs	Sqr	Lin	Log	Scl	1.25	1.25^2	1.25^3	PSNR	SSIM	LPIPS
Right Eye	M_1	7.34	4636	0.07	1.15	10.4	0.24	0.38	2.74	2.93	2.97	28.39	0.63	2.905
	M_2	7.14	4511	0.07	1.15	10.1	0.24	0.37	2.75	2.93	2.97	28.81	0.63	2.929
	M_3	6.57	4224	0.07	0.95	9.53	0.23	0.38	2.78	2.94	2.98	29.16	0.65	2.844
Left Eye	M_1	7.53	4664	0.07	1.13	10.5	0.25	0.41	2.75	2.93	2.98	28.29	0.62	2.926
	M_2	7.59	4775	0.08	1.22	10.7	0.26	0.42	2.74	2.93	2.97	28.21	0.60	3.063
	M_3	7.04	4457	0.07	1.07	10.1	0.25	0.42	2.77	2.93	2.97	28.77	0.61	3.089
Nose	M_1	7.52	4433	0.07	0.96	10.0	0.46	0.79	2.82	2.96	2.99	28.86	0.57	3.115
	M_2	6.21	3715	0.06	0.66	8.38	0.43	0.74	2.89	2.98	2.99	30.17	0.60	2.894
	M_3	7.93	4744	0.07	1.16	10.7	0.55	0.95	2.77	2.94	2.97	28.67	0.54	3.257
Mouth	M_1	7.63	4571	0.07	0.99	10.3	0.18	0.28	2.82	2.96	2.99	28.44	0.56	3.312
	M_2	7.01	4299	0.06	0.86	9.70	0.17	0.27	2.85	2.97	2.99	28.96	0.58	3.211
	M_3	6.71	4106	0.06	0.81	9.26	0.16	0.26	2.86	2.97	2.99	29.45	0.59	3.133
All Pacthes	M_1	7.50	4576	**0.07**	1.06	10.3	0.28	0.46	2.78	**2.95**	**2.98**	28.50	0.60	3.065
	M_2	**6.99**	**4325**	**0.07**	**0.97**	**9.75**	**0.27**	**0.45**	**2.81**	**2.95**	**2.98**	**29.03**	**0.61**	**3.024**
	M_3	7.06	4383	**0.07**	1.00	9.88	0.30	0.50	2.79	**2.95**	**2.98**	29.01	0.60	3.081

three δ-metrics, *i.e.* the percentage of pixels under a given threshold. In our analysis, we include also the *Peak Signal-to-Noise Ratio* (PSNR), that estimates the logarithmic level of noise defined as:

$$\text{PSNR} = 10 \cdot \log_{10} \frac{\max I}{L_2} \tag{12}$$

where $\max I$ is the maximum possible value of the ground truth image I and the generated image (in our experiments, we set $\max I = 255$). We also use the *Structural Similarity* (SSIM), which estimates the perceived visual similarity of two images, defined as:

$$\text{SSIM}(w_{1,2}) = \frac{(2\mu_{w_1}\mu_{w_2} + c_1)(2\sigma_{w_{1,2}} + c_2)}{(\mu^2_{w_1} + \mu^2_{w_2} + c_1)(\sigma^2_{w_1} + \sigma^2_{w_2} + c_2)} \tag{13}$$

Given two windows w_1, w_2 of equal size, $\mu_{w_{1,2}}$, $\sigma_{w_{1,2}}$ are the mean and variance of w_1, w_2 while $c_{1,2}$ are used to stabilize the division. Further details are reported in [47]. We believe that these metrics can capture the differences in pixels between the real and the retouched images.

Moreover, we exploit the *Learn Perceptual Image Patch Similarity* (LPIPS) [48],[2] based on deep neural network activations (in our tests we use VGG-16 [49] trained on *Imagenet* [50] dataset), to estimate the visual quality of generated images. In the work of Zhang *et al.* [48], it has been shown that LPIPS metrics is strictly related to human judgment.

1) ATTENTION MAP ANALYSIS

As a first step in the performance analysis of the framework, we investigate the impact of the use of different types of Attention Maps. Specifically, we compare the three different maps, referred to as M_1, M_2 and M_3, computed as detailed in Section III-B and depicted in Figure 3.

Results are reported in Table 2, in which, for the sake of comprehension, the first lines show the quantitative results obtained on each single patch type, and the final row contains the average values computed on all patch types. Average values indicated that the second type of the map, M_2, represents the best choice for the framework, even though for certain metrics the difference of values is limited. However, the higher PSNR value reveals that the output tends to be less affected by noise. Furthermore, L_1 and L_2 metrics, computed on the distance between pixel values, indicate that the framework is able to retouch patches in specific areas. Starting from these considerations and results, we selected the hyper-parameters empirically found in this experiment and we used M_2 as the attention map in all the following experiments.

2) ABLATION STUDY

As the second step, we directly analyze the impact of using M_2 in input through an Ablation Study. In particular, we compare the performance of the presented framework with and without the use of the Attention Map. Results are shown in Table 3, in which we report, as in the previous case, the pixel-wise metrics computed on each patch type and, at the bottom of the table, the average evaluation of the collected values.

We observe that the impact of the Attention Map is largely positive since all the pixel-wise metrics show a substantial improvement. In particular, the PSNR metric value (in logarithmic scale) reveals that the use of Attention Maps can significantly reduce the amount of noise in generated images, *i.e.* the conditional GAN is focused to change only a limited amount of pixels in the patch. Also, the distance

[2] https://github.com/richzhang/PerceptualSimilarity

TABLE 3. Pixel-wise metrics computed on the FRGC$_S$ dataset for the ablation study of the proposed framework. Specifically, the values reflect the impact of the use of attention maps in the input of the conditional GAN. On the top, arrows indicate the positive changing direction, in which better performance corresponds to a value variation that follows the arrow.

Patch Type	Method	Norm ↓		Difference ↓		RMSE ↓			δ-metrics ↑			Indexes ↑		Perc. ↓
		L_1	L_2	Abs	Sqr	Lin	Log	Scl	1.25	1.25^2	1.25^3	PSNR	SSIM	LPIPS
Right Eye	No Map	7.84	5113	0.08	1.26	11.5	0.25	0.40	2.72	2.92	2.97	27.15	0.61	3.002
	M_2	7.14	4511	0.07	1.15	10.1	0.24	0.37	2.75	2.93	2.97	28.81	0.63	2.929
Left Eye	No Map	7.62	4792	0.08	1.16	10.8	0.25	0.42	2.74	2.93	2.97	27.96	0.62	3.051
	M_2	7.59	4775	0.08	1.22	10.7	0.26	0.42	2.74	2.93	2.97	28.21	0.60	3.063
Nose	No Map	8.77	5272	0.08	1.29	11.9	0.49	0.84	2.76	2.94	2.98	27.11	0.54	3.054
	M_2	6.21	3715	0.06	0.66	8.38	0.43	0.74	2.89	2.98	2.99	30.17	0.60	2.894
Mouth	No Map	8.24	4895	0.07	1.17	11.0	0.19	0.30	2.77	2.95	2.98	27.95	0.55	3.200
	M_2	7.01	4299	0.06	0.86	9.70	0.17	0.27	2.85	2.97	2.99	28.96	0.58	3.211
All Pacthes	No Map	8.12	5018	0.08	1.22	11.32	0.30	0.49	2.75	2.94	2.98	27.54	0.58	3.077
	M_2	**6.99**	**4325**	**0.07**	**0.97**	**9.75**	**0.27**	**0.45**	**2.81**	**2.95**	**2.98**	**29.03**	**0.61**	**3.024**

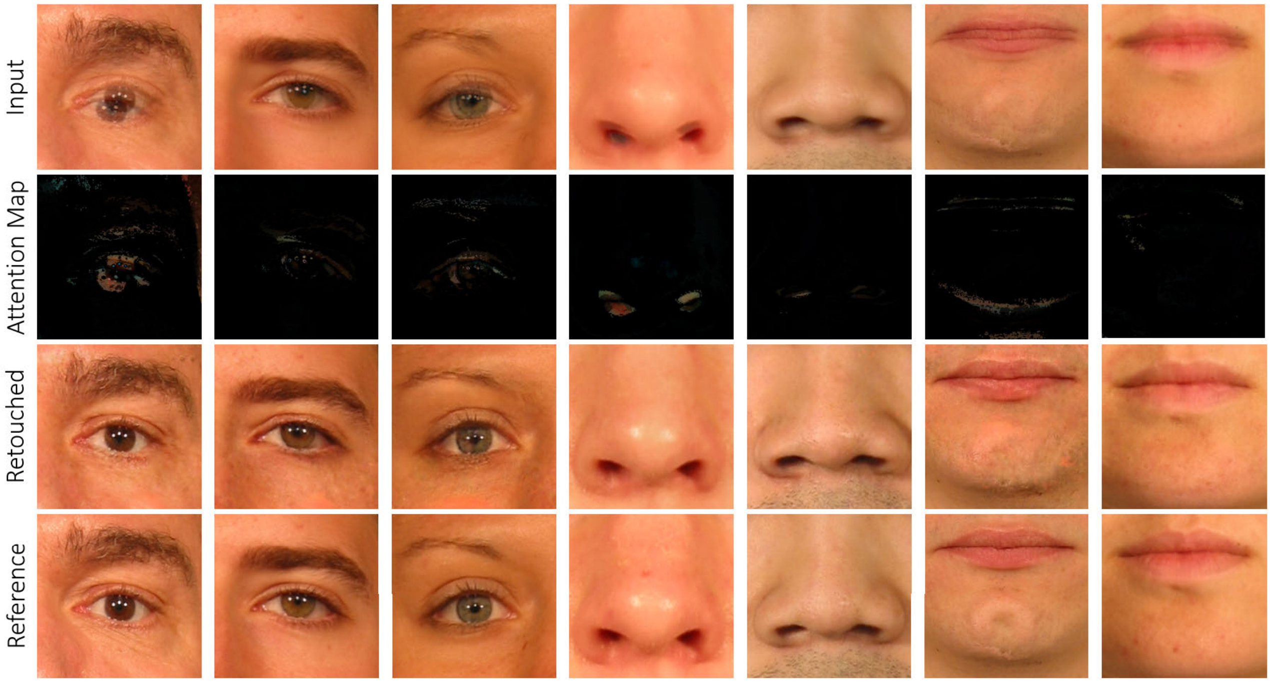


FIGURE 6. Example of generated images on the FRGC$_S$ dataset. In the first line, original images with artifacts are reported. Then, the M_2 attention maps, the generated and reference images are shown in the following lines. (best zoomed on screen).

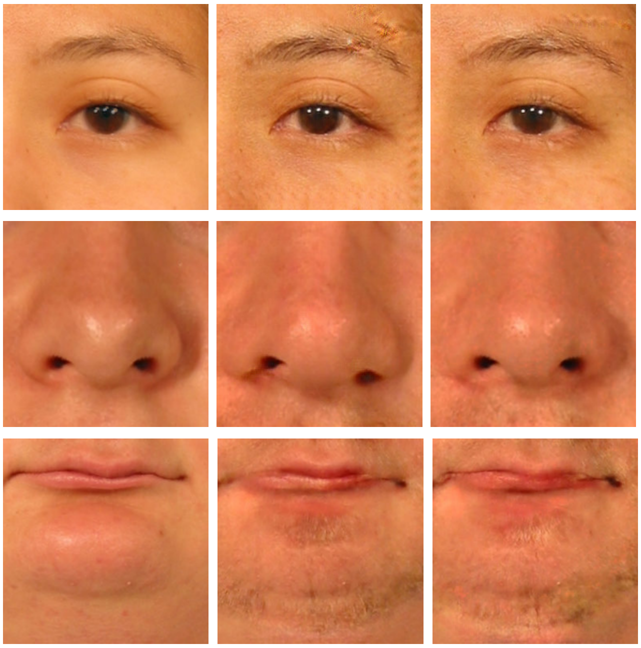
metrics, ranging from L_1 to the RMSE confirm that the generated patches have superior quality. Moreover, some qualitative results are reported in Figure 7, in which we observe that the quantitative results reflect the quality of the generated images. Indeed, the effects of the use of Attention Maps in the input are visible, in the three different patches. We note that generally, the presence of artifacts is limited in space, since visible artifacts are less scattered along the contours of the image, such as in the eye patch of the first line. Moreover, we note that thanks to the Attention Map the retouching procedure is focused also on small details, as shown in the right naris and the lips of the second and third rows.

3) EXTERNAL COMPARISON

Finally, we carried out a comparison between the proposed approach and the most similar work from the recent literature, the style transfer-based approach [17] previously introduced and described in Section II. Quantitative results are reported in Table 4. In the first line, referred to as "baseline", we show the values obtained comparing reference images with initial patches, *i.e.* the patches not automatically retouched by the proposed framework: we believe these numbers provide a baseline and increase the understanding of the real improvement introduced by the retouching operation. As shown, our approach generally overcomes the baseline and the literature competitor in the large majority of reported metrics. L_1 and L_2

TABLE 4. Experimental results of pixel-wise metrics computed on generated images from FRGC$_S$ dataset. Further details about metrics are reported in Section IV-B and [43]. On the top, arrows indicate the positive changing direction, in which better performance corresponds to a value variation that follows the arrow.

Patch Type	Method	Norm ↓		Difference ↓		RMSE ↓			δ-metrics ↑			Indexes ↑		Perc. ↓
		L_1	L_2	Abs	Sqr	Lin	Log	Scl	1.25	1.25^2	1.25^3	PSNR	SSIM	LPIPS
Right Eye	Baseline	7.74	4889	0.07	1.18	11.1	0.24	0.39	2.75	2.93	2.97	27.65	0.63	3.003
	[17]	8.26	5557	0.08	1.63	12.5	0.16	0.25	2.70	2.90	2.96	26.86	0.50	3.121
	Ours	7.14	4511	0.07	1.15	10.1	0.24	0.37	2.75	2.93	2.97	28.81	0.76	2.929
Left Eye	Baseline	7.86	4966	0.07	1.20	11.2	0.25	0.41	2.75	2.93	2.97	27.57	0.60	3.025
	[17]	8.12	5487	0.08	1.55	12.3	0.16	0.25	2.72	2.91	2.96	26.94	0.50	3.100
	Ours	7.59	4775	0.08	1.22	10.7	0.26	0.42	2.74	2.93	2.97	28.21	0.75	3.063
Nose	Baseline	7.87	4744	0.08	1.13	10.7	0.47	0.81	2.78	2.94	2.97	27.97	0.60	3.991
	[17]	8.39	4611	0.13	1.07	10.7	0.29	0.48	2.37	2.75	2.88	21.95	0.60	3.211
	Ours	6.21	3715	0.06	0.66	8.38	0.43	0.74	2.89	2.98	2.99	30.17	0.69	2.894
Mouth	Baseline	7.95	4298	0.06	0.91	9.69	0.18	0.30	2.82	2.96	2.99	28.84	0.58	3.509
	[17]	7.64	4393	0.09	1.29	10.4	0.17	0.27	2.63	2.88	2.95	28.38	0.58	3.342
	Ours	7.01	4299	0.06	0.86	9.70	0.17	0.27	2.85	2.97	2.99	28.96	0.69	3.211
All Pacthes	Baseline	7.85	4724	**0.07**	1.11	10.7	0.29	0.48	2.78	2.94	**2.98**	27.51	0.61	3.382
	[17]	8.10	5012	0.10	1.39	11.5	**0.20**	**0.31**	2.61	2.86	2.94	24.78	0.55	3.194
	Ours	**6.99**	**4325**	**0.07**	**0.97**	**9.75**	0.27	0.45	**2.81**	**2.95**	**2.98**	**29.03**	**0.72**	**3.024**



(a) Input (b) Without Attention (c) With Attention

FIGURE 7. Visual example on the effects of introducing an attention map in the generative process. In (a) is reported the reference image, the output of the framework without and with the use of the Attention Map is shown in (b) and (c), respectively. Generally, in (c) artifacts related to the cGAN generation, are less visible or more limited in space.

metrics reveal that our framework is able to accurately adjust the intensity of pixel values. It is also important to note that the PSNR, which reveals the presence of noise in images and is expressed in a logarithmic scale, is largely better. Also, the perceptual similarity metric (LPIPS) confirms the good visual results obtained, together with the SSIM. We note the style transfer method [17] does not perform well on this dataset: probably, pixel-wise metrics penalize the limited size of generated images (only 224 × 224 for the whole face), and the salt-and-pepper noise visible in them. Moreover, this method requires a significant amount of time to be evaluated on a single input image due to the optimization process. Qualitative results of the proposed method are shown in Figure 6: in the first line, there are the input patches with artifacts along with the related Attention Maps on the second line. Then, the retouched and reference images are depicted, respectively. The visual results confirm the effectiveness of the proposed approach in terms of artifact removal. The modifications introduced are mainly focused on the regions highlighted by the Attention Map, even though small GAN-generated artificial details are rarely visible.

C. FRAMEWORK ANALYSIS THROUGH FRGC$_M$

In this section, we test our system with morphed images contained in the FRGC$_M$ set. In this setting, differently from FRGC$_s$ dataset, a ground truth reference image is missing; however, FRGC$_M$ allows the testing of the framework with real morphed images. Several qualitative results are depicted in Figure 8, in which the last line shows the retouched patches computed starting from the morphed ones in the first row. As shown, the overall visual quality of the generated images is adequate, especially in comparison with the input ones: indeed, the proposed system is effectively able to retouch artifacts in terms of ghosts, blurred areas, and different levels of color equalization. The improvement is clear for each patch type, the retouching result is visible in specific areas such as near the contour of the iris and the nostrils, and the internal part of the mouth lips. In the third row of the table, we report also the difference between the final retouched patch and the input image: in this manner, we are able to visually highlight the areas retouched by the cGAN and to compare this result with the Attention Map. This visualization confirms that cGAN acts on very limited areas modifying pixels in specific

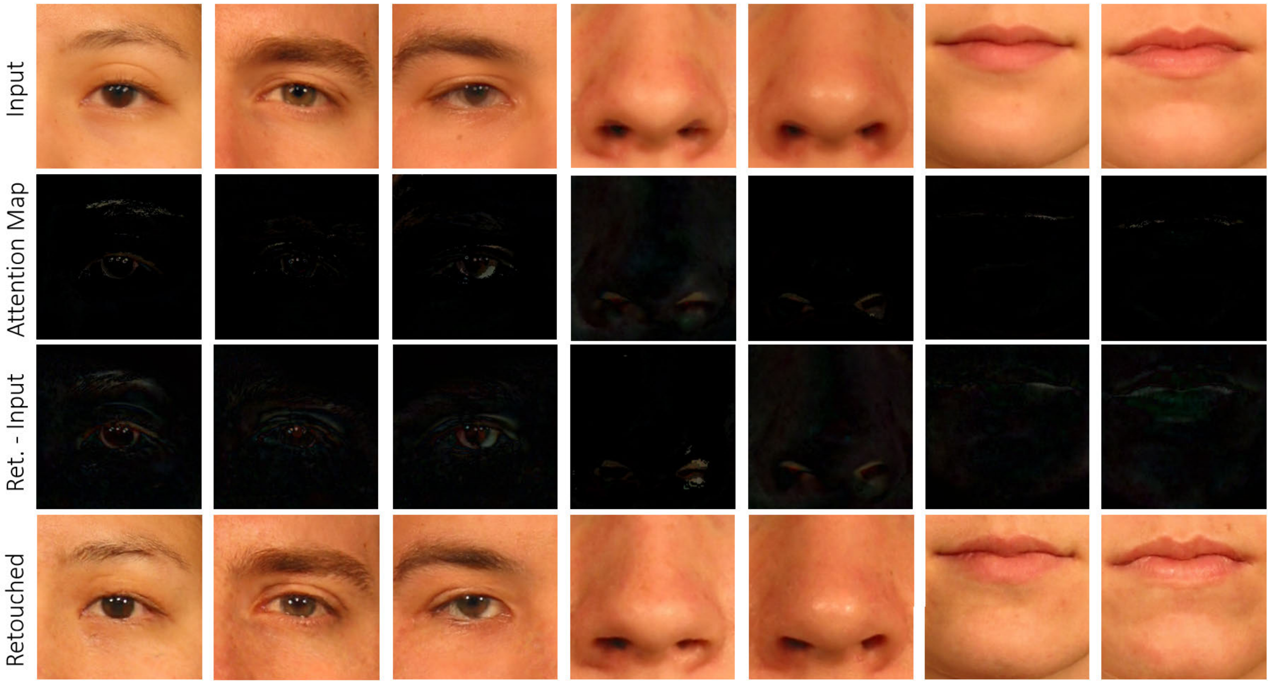


FIGURE 8. Example of generated images on the FRGC$_M$ dataset. In the first line, original images with artifacts are reported. The M_2 attention map is reported in the second line, while the final retouched patches are reported in the last row. In the third line, we show the difference between images in line 4 and line 1, in order to highlight the areas changed by the proposed framework to be compared with the related attention map. (best zoomed on screen).

positions suggested by the Attention Map, even though a small amount of background noise is present. This property is very important since the small changes introduced by the retouching process do not significantly alter the image content, as discussed in the next section.

1) OVERALL VISUAL QUALITY

The overall quality of retouched images can be appreciated in Figure 9, which reports the whole faces obtained after the automatic retouch and the weighted blending procedure. The images refer to 10 different morphing examples, 5 men (left) and 5 women (right). A visual analysis of the results obtained confirms the efficacy of the proposed approach in removing the typical morphing artifacts. The general aspect is preserved, the main facial features remain unaltered and the intervention is limited to small image areas. A significant improvement can be observed in the eye region, where the iris artifacts are successfully removed, as well as the double reflections in the pupils that usually characterize morphed images. The result obtained is satisfactory and quite realistic even for subjects wearing eyeglasses (see the fourth row): the eye definition and sharpness are improved without causing anomalous distortions to the glasses frame that could reveal an image alteration. Analogous results are obtained for the nose and mouth region where double edges in the nostrils or lips are effectively removed. Some details worth of attention can be observed in the mouth region where also major defects, deriving for instance from the mouth open in one of the two parent images (see last two rows), are successfully addressed; in general, in these cases, the proposed approach tends to regularize anomalies by, for instance, “closing” the mouth, as shown in the examples.

2) IDENTITY PRESERVATION

One important requirement for automated post-processing is the preservation of the identity associated with the morphed image, meaning that if a morphed image can be successfully matched to both parent subjects before retouching, this property should persist even after automated artifact removal. To analyze this phenomenon, we adopt here the *Mated Morph Presentation Match Rate* (MMPMR) metric, proposed in [4] as a measure of the vulnerability of FRSs. MMPMR represents the proportion of morphed images that can be successfully matched with both parent subjects and is defined as:

$$\text{MMPMR} = \frac{1}{M}\sum_{m=1}^{M}[min_{n=1\ldots N_m} S^n_{I_m}] > \tau \qquad (14)$$

where, M is the number of morphed images in the dataset, N_m is the number of parent subjects for a specific morphed image I_m ($N_m = 2$ here), $S^n_{I_m}$ is the comparison score for the morph I_m of subject n and τ being the threshold of FRS at a chosen *False Acceptance Rate* (FAR). For our experiments, we used two FRSs: i) a commercial SDK, VeriLook (version 12) by Neurotechnology, and ii) ArcFace [51], an open-source deep-learning-based solution. For the face verification

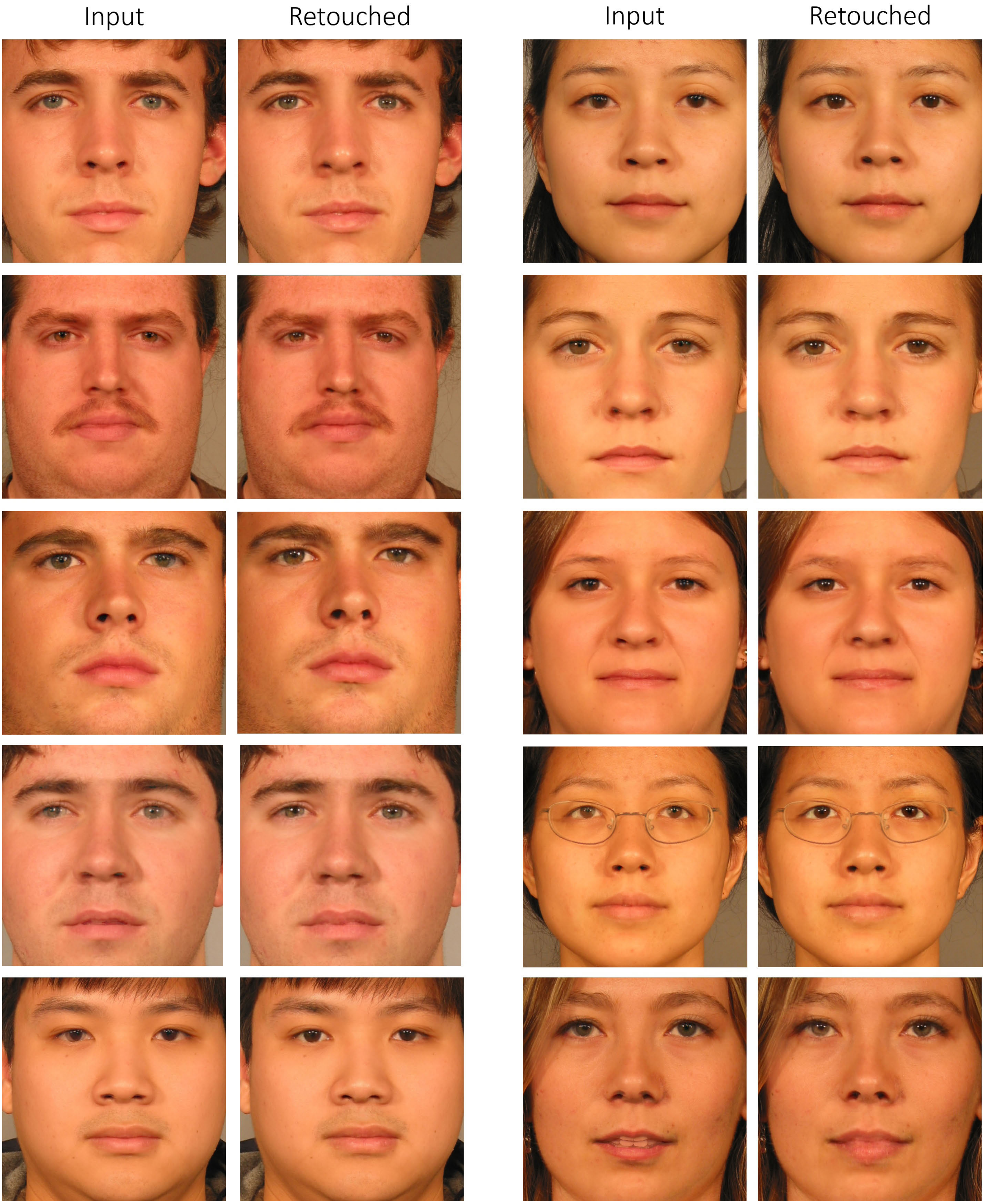
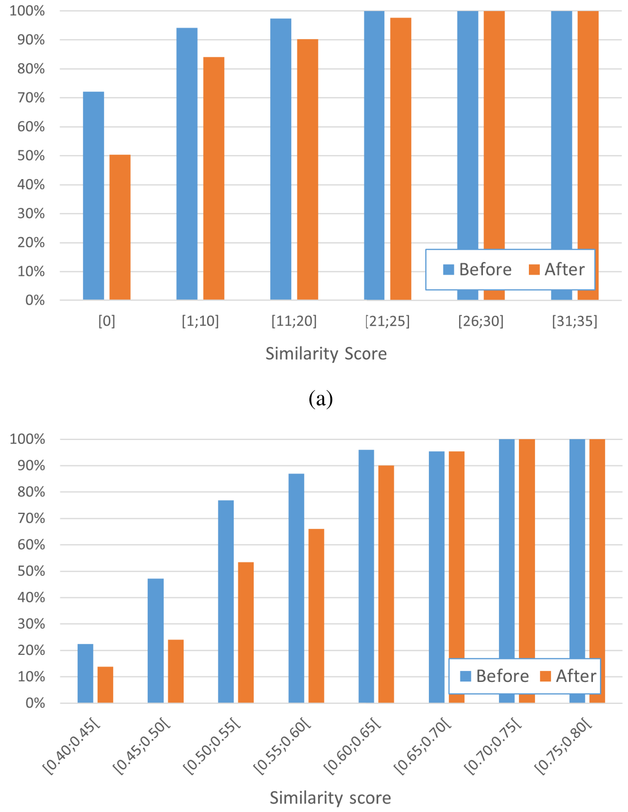


FIGURE 9. The final output of the proposed framework, that consists of whole faces automatically retouched by a weighted blending in the starting morphed face. For each column, on the left, the initial morphed face with artifacts is shown, while the retouched face is reported on the right. (best zoomed on screen).



(a)

(b)

FIGURE 10. MMPMR values measured in the identity preservation test as a function of the similarity score between the two parent images used for morphing; the results are reported for two FRS: (a) VeriLook v12 by Neurotechnology (commercial) and (b) ArcFace (open source).

experiments, each morphed image m is compared against a test image of each parent subject (different from the one used for morphing creation). Following the Frontex guidelines for face verification at ABC gates, the threshold has been fixed for both SDKs in order to operate at a FAR of 0.1%; while for VeriLook the threshold to be used is provided by the SDK, for ArcFace we established it on the basis of an internal test including a number of impostor attempts.

Figure 10 reports the MMPMR values measured for the two SDKs before and after automated artifact removal. It is generally recognized that high similarity between the two parent subjects used for morphing noticeably increases the effectiveness of the morphing attack; for this reason, the graphs report the MMPMR values for discretized values of similarity between the two contributing images. The graphs confirm a common trend for the two SDKs and clearly show that, for high similarity scores, the automated retouching does not affect MMPMR, thus confirming that the morphed identity is preserved. For lower similarity scores, a reduction of MMPMR is observed after retouching; this phenomenon is not unexpected, since generally in this case the chances of success are quite limited, the matching score is low and even small image modifications often cause it to drop below the established matching threshold. According to our direct experience, the same behaviour is observed when the retouching is manually executed by human experts and sometimes even the alteration of apparently insignificant details has an impact on the FRS similarity score.

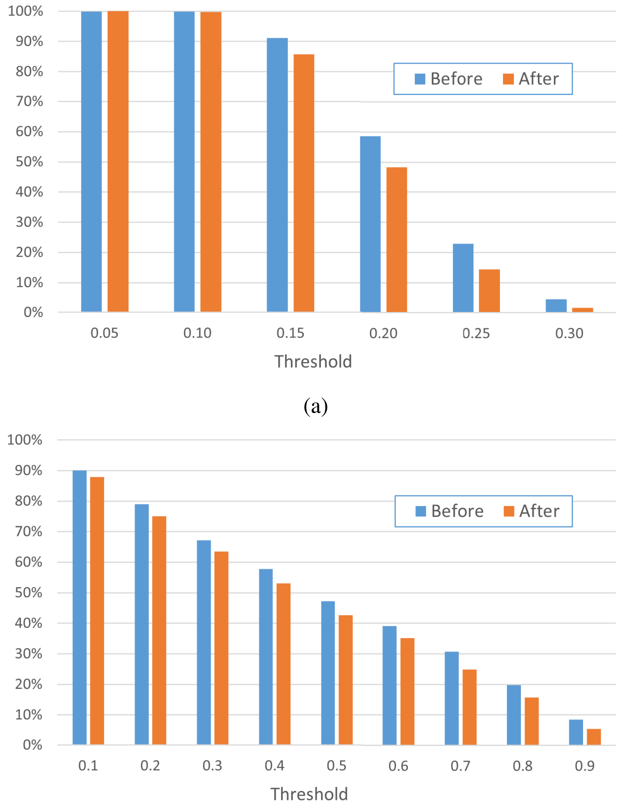
Overall we can conclude that, when the morphed image has high chances of success, i.e. it has been generated from quite similar subjects, the proposed automated retouching doesn't affect the probability of fooling the FRS.

3) MORPHING ATTACK DETECTION

In this section, we investigate whether generated images are able to fool MAD algorithms and whether they can be used for data augmentation to improve the training procedure. Specifically, in the following tests, we address the Single image-based MAD (S-MAD) task, since it perfectly suits the focus of our work: during the testing phase, for each input image, an S-MAD algorithm outputs a value in the range of [0, 1], representing how confident the method is that the input image is morphed (0 means that image does not contain a morph while 1 reveals certainty that the image contains a morph). We use the images in FRGC$_M$ as the test set. In this case, all images are labeled as morphed.

Firstly, we assess if the quality of retouched images is suitable to fool an algorithm developed for the S-MAD task: indeed, the goal of the proposed framework is to automatically retouch morphed images, leading to a high-quality result comparable with the time-consuming manual retouch. Two different methods available in the literature are selected: the first one is based on a Machine Learning approach and it is described in [30] and summarized in Section II. This method has been selected since, currently, it has state-of-the-art performance on public evaluation platforms [52]. The second method is based on Deep Neural Network and it is inspired by the recent work described in [32]: a ResNet-18 [53] architecture, pre-trained on the Imagenet dataset [50], is fine-tuned on the PMDB [35] dataset. Results are reported in two histograms in Figure 11. On the x-axis we report the thresholds used to classify if an input image is morphed or not, while in the y-axis we report the percentage of images correctly detected as morphed. We note that the proposed framework effectively introduces an improvement of quality in tested morphed images since orange bars in Figure 11a reveal a lower morphing detection percentage. Therefore, retouched images are able to fool the S-MAD detector better than the original morphed images. We report similar observation also for the deep learning-based method in Figure 11b, since also in this case the percentages of detected morphed images are lower.

Secondly, we investigated if the retouched images can be useful to more robustly train a MAD algorithm. In other words, we test the possibility to use the proposed framework to augment the quantity and quality of training data exploitable by MAD algorithms, especially if deep learning-based. Therefore, we adopt, as in the previous case, a ResNet-18 [53] architecture, pre-trained on Imagenet dataset [50]. Then, we test its performance on the MorphDB [35] dataset



(a)

(b)

FIGURE 11. Results of the S-MAD algorithms computed on the retouched images from FRGC$_M$ dataset. (a) is the machine learning-based method described in [30], while (b) is a deep learning-based introduced in [32].

that contains 100 manually retouched high-quality morphed images. Specifically, we compare two training procedures. In the first one, the deep neural network is trained only on the PMDB dataset, while in the second case the network is trained on the images contained in the PMDB dataset merged with the retouched ones in the FRGC$_M$ dataset. Results are reported in Figure 12 in the form of *Detection Error Trade-off* (DET) curve. In red, the curve of the system trained with only the PMDB data, while in blue the curve of the framework trained with the merged data. Although the absolute performance of the S-MAD approach can be further improved, we note that the use of retouched images effectively increases the efficacy of the training procedure, leading to more accurate results. Probably, the use of retouched images helps to prevent overfitting phenomena.

4) HUMAN OBSERVER ANALYSIS

In the final part of our work, we describe the qualitative assessment made by human observers regarding the retouched morphed images. To collect human evaluations, we developed a web page where the images are shown in a grid and the user has to indicate what images are altered. We do not explicitly indicate the type of artifacts present in the images, since this will instill a bias in the participant choices. The images are displayed in a grid and for each of them, the users have to indicate if they think the image was altered in any way. This setting is similar to a single morphing attack detection (S-MAD) scenario, where no reference image is present for comparison. Note that the aim of this test is to qualitatively assess the performance of our retouching algorithm, in terms of the realness of the generated images.

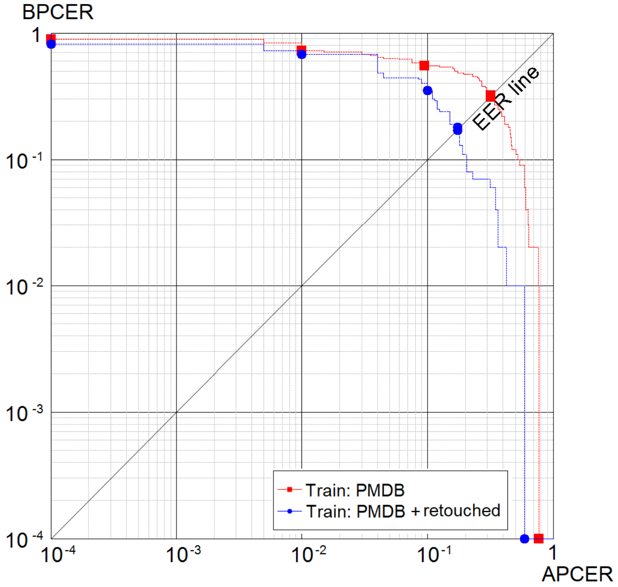


FIGURE 12. *Detection Error Trade-off* (DET) curve computed on MorphDB dataset [35]. In red, the curve of the model trained only with PMDB data [35], while in blue the values of the model trained with PMDB and retouched images from FRGC$_M$ dataset.

We selected a total of 75 images from 5 different groups: bona fide (unaltered, not morphed, images), morphed images without retouching, morphed images after retouching, images retouched through the approach described in [17], and the ones generated by MIPGAN [7]. For each group, we selected 15 images, with a near equal number of male and female subjects. The images are displayed on five pages with 15 images each (in order to not overload and stress the evaluator with many images at the same time). Each page contains 3 images for each group and a similar number of female and male subjects. The placement of the images in the pages is totally random and no predetermined scheme is used. We do not impose any time limit to the evaluation, but we encouraged the participants to spend a maximum of one minute per page, so approximately four seconds per image, in order to reproduce a control at ABC gates. A screenshot of a part of the web page developed is shown in Figure 14.

A total of 57 people participated in the evaluation, both experts/experienced people in the field of face morphing, and inexperienced people. The group of experienced people is composed of researchers and experts in the field of face morphing, while the inexperienced group is composed of students, computer science professionals, and even people not related to computer vision or computer science.

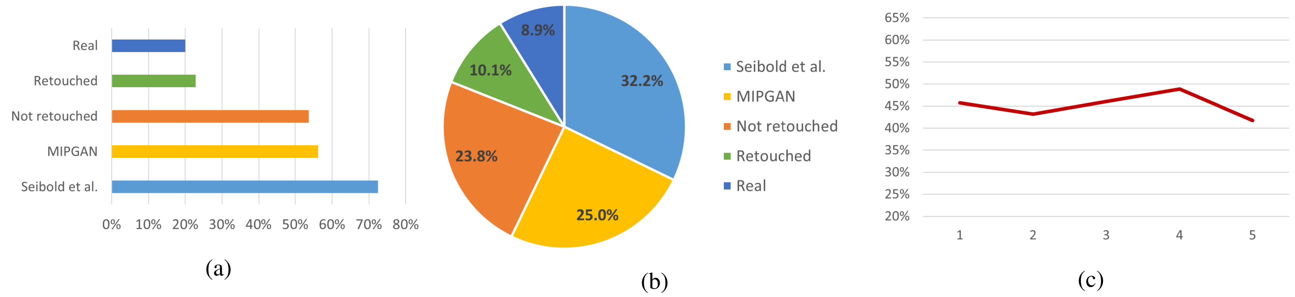


FIGURE 13. Results of the human evaluation analysis. (a): percentages of images of each category indicated as altered; (b): percentages of images of each category indicated as altered, computed over the final set of images selected by participants; (c): percentages of images indicated as altered for every page of the evaluation form. Each page includes the same number of images for each category, so the same number of altered images. The number of indicated images should be constant throughout the test to indicate a constant level of attention of the participants throughout the test.

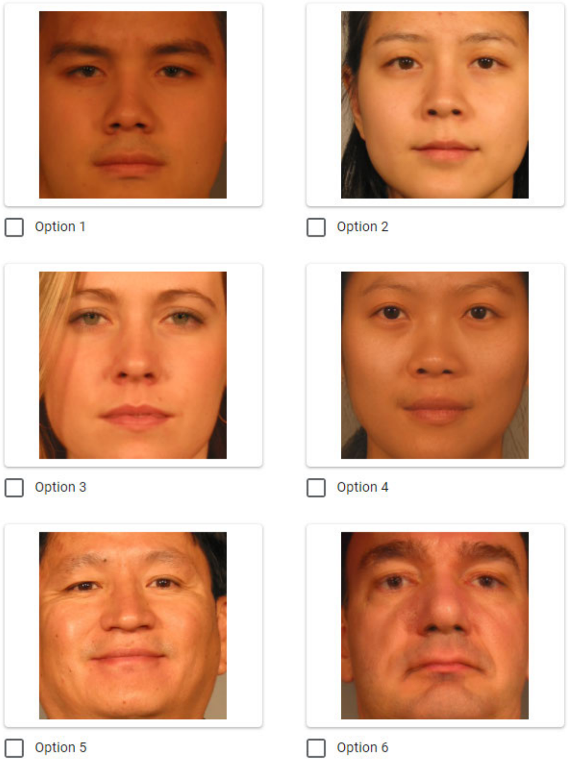


FIGURE 14. A screenshot of a part of the web page used for human observer analysis. The images from the five groups are displayed randomly in a grid, for a total of 15 options, and the participants have to select the ones they believe are altered checking the corresponding check box under the image. In the image, "Option 1" and "Option 3" show a morphed image before the retouch, "Option 2" a retouched image, "Option 4" a MIPGAN image, "Option 5" a real (bona fide) image and "Option 6" the output of [17].

The results are shown in Figure 13. The effectiveness of the proposed retouch method is confirmed by the results of the human observer analysis. With respect to the number of images of each category, morphed images not retouched are indicated as altered 23.8% times, but after retouching the number drops to 10.1%, with an improvement of 13.7% in terms of missed detections. This percentage is similar to the evaluation of real (bona fide) images, which are indicated as altered 8.9% times. This can be explained since some of the chosen real images are slightly blurred or the person does not directly watch into the camera. Nevertheless, these results indicate that the retouched images have a visual quality similar to the real ones. In the case of MIPGAN [7] ("MIPGAN" in the graphs) and style transfer-based work [17] (referred to as "Seibold *et al.*" in the graphs), the examined people indicated 25.0% and 32.2% altered images respectively. It is important to note that MIPGAN is not oriented to the retouching of artifacts, but it is a GAN-based morphing generator. Therefore, a direct comparison between our framework and MIPGAN is not fair and it is out of the scope of this paper. However, it is interesting including also those images in the test in order to understand the impact of GAN-related artifacts on human morphing detection capabilities. Moreover, in case of "Seibold *et al.*", the method was initially conceived for face enhancement, especially for correcting face imperfections, and it was not designed to correct the typical morphing artifacts. Nevertheless, a comparison with this method may provide a solid baseline w.r.t. state-of-the-art face enhancement algorithms. Indeed, the MIPGAN images appear quite blurred and the skin and facial hairs are smoothed. In general, for the participants identifying morphed images produced by MIPGAN and by landmark-based morphing algorithms has the same difficulty, since the percentage of images selected as altered is similar for both the categories. Style you Face Morph [17] presents the worst results, and it seems quite easy for the humans involved in this experiment to spot the images retouched with it. This can be explained by the fact that the method is mainly focused on texture artifacts and the smoothing effect often present in the morphed images.

To further validate the results we conducted an analysis to understand if the attention of the participants remains stable during the entire duration of the test or it decays in the final part of the experiment. Thus, we examined the number of indicated altered images for every page. As discussed before, there are five pages, each of them composed of 15 images. The same number of images for each category is present on every page, so we expect that the number of images indicated as altered remains constant throughout the pages. A decaying

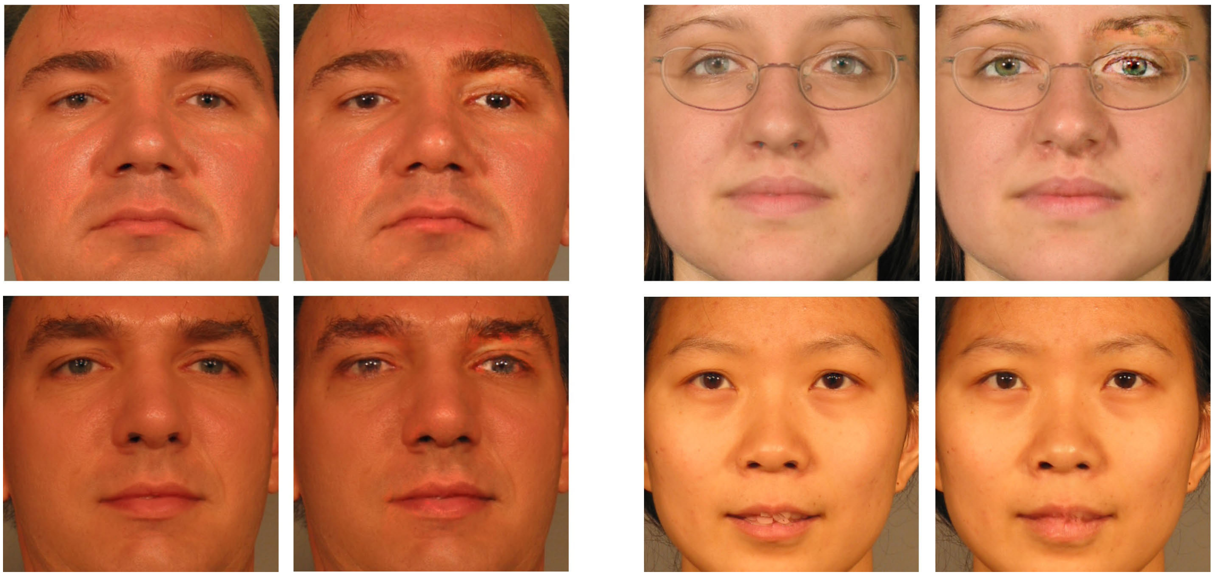


FIGURE 15. Some examples of failure cases, in which it is possible to note the presence of typical texture artifacts introduced by GANs or not optimal retouches in some regions.

number of selected images in the final pages may indicate a drop in the attention of the participants. The results are reported in Figure 13c. As it can be seen the percentage of altered images indicated by the participants on every page is almost the same, and there is no drop in the number as the evaluation progresses. The number of indicated images slightly fluctuates between pages, but this can be explained by the different difficulties of the included images. This further analysis confirms that the level of attention remained stable throughout the test, validating the results and the methodology used to assess human observer analysis.

V. CONCLUSION

In this paper, a system to automatically retouch visual artifacts generated by a landmark-based morphing process is presented. The system, based on a Conditional GAN, is fed with the concatenation of an RGB image and the related Attention Map derived from the two warped subjects involved in the morphing process. The final retouched patches are then blended in the original morphed face. The quantitative and qualitative results are very encouraging. Indeed, a variety of investigations have been conducted in order to assess the visual quality of retouched patches, identity preservation, and the ability to fool both human observers and MAD algorithms. Moreover, the utility of the generated images for data augmentation in MAD training has been proved as well. Therefore, this work can be one of the first successful investigations on the use of a GAN-based approach to automatically retouch morphed images.

Although the overall results are satisfactory, some failure cases have been observed in our experiments as shown in Figure 15. The main causes of errors can generally be identified in relation to particular lighting conditions, suggesting the adoption of a more effective image pre-processing to reduce this phenomenon; however, it is worth mentioning that this situation is unlikely to happen in a real case where the images used for morphing are well controlled and fully ISO/ICAO compliant. Other cases refer to specific image features (e.g. eyeglasses or open mouth) that are underrepresented in the training set and therefore more difficult to address.

A variety of future work can be planned. The final weighted blending procedure can be improved, for instance, by exploiting a deep learning-based method, following the recent in-painting works available in the literature. Furthermore, the attention paradigm can be implemented not only in input through the use of Attention Maps, but, for instance, can be injected inside the deep neural network, through the use of attentional mechanism or transformer-based architectures. Again, the working spatial resolution of patches can be increased through the use of specific architectures and appropriate Graphical Processing Units (GPUs), which can lead to overcoming the patch-based approach and enable processing the whole face at once. Finally, the patch-based approach in input can be overcome in order to avoid limitations related to the fixed size of the patches and the different statistical characteristics of pixels that can divert from the different face areas.

ACKNOWLEDGMENT

In addition, the authors would like to thank Dr. Leonardo Odorisio and Ing. Andrea De Maria for the insightful discussions on the subject. Disclaimer: this text reflects only the author's views, and the commission is not liable for any use that may be made of the information contained therein.

REFERENCES

[1] *Face Recognition Grand Challenge (FRGC)*. Accessed: Mar. 22, 2021. [Online]. Available: https://www.nist.gov/programs-projects/face-recognition-grand-challenge-frgc

[2] M. Ferrara, A. Franco, and D. Maltoni, "The magic passport," in *Proc. IEEE Int. Joint Conf. Biometrics*, Sep. 2014, pp. 1–7.

[3] U. Scherhag, C. Rathgeb, J. Merkle, R. Breithaupt, and C. Busch, "Face recognition systems under morphing attacks: A survey," *IEEE Access*, vol. 7, pp. 23012–23026, 2019.

[4] U. Scherhag, A. Nautsch, C. Rathgeb, M. Gomez-Barrero, R. N. J. Veldhuis, L. Spreeuwers, M. Schils, D. Maltoni, P. Grother, S. Marcel, R. Breithaupt, R. Ramachandra, and C. Busch, "Biometric systems under morphing attacks: Assessment of morphing techniques and vulnerability reporting," in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Sep. 2017, pp. 1–7.

[5] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 2. Cambridge, MA, USA: MIT Press, 2014, pp. 2672–2680.

[6] N. Damer, A. M. Saladie, A. Braun, and A. Kuijper, "MorGAN: Recognition vulnerability and attack detectability of face morphing attacks created by generative adversarial network," in *Proc. IEEE 9th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Oct. 2018, pp. 1–10.

[7] H. Zhang, S. Venkatesh, R. Ramachandra, K. Raja, N. Damer, and C. Busch, "Mipgan—Generating robust and high quality morph attacks using identity prior driven GAN," 2020, *arXiv:2009.01729*. [Online]. Available: https://arxiv.org/abs/2009.01729

[8] L. Chai, D. Bau, S.-N. Lim, and P. Isola, "What makes fake images detectable? Understanding properties that generalize," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 103–120.

[9] 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. (CVPR)*, Jul. 2017, pp. 1125–1134.

[10] K. Raja *et al.*, "Morphing attack detection—Database, evaluation platform and benchmarking," 2020, *arXiv:2006.06458*. [Online]. Available: http://arxiv.org/abs/2006.06458

[11] A. Shafaei, J. J. Little, and M. Schmidt, "Autoretouch: Automatic professional face retouching," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 990–998.

[12] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. Efros, "Detecting photoshopped faces by scripting photoshop," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10072–10081.

[13] L. Bondi, S. Lameri, D. Guera, P. Bestagini, E. J. Delp, and S. Tubaro, "Tampering detection and localization through clustering of camera-based CNN features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1855–1864.

[14] J. Li, X. Li, B. Yang, and X. Sun, "Segmentation-based image copy-move forgery detection scheme," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 3, pp. 507–518, Mar. 2015.

[15] A. C. Popescu and H. Farid, "Exposing digital forgeries by detecting traces of resampling," *IEEE Trans. Signal Process.*, vol. 53, no. 2, pp. 758–767, Feb. 2005.

[16] J. F. O'Brien and H. Farid, "Exposing photo manipulation with inconsistent reflections," *ACM Trans. Graph.*, vol. 31, no. 1, pp. 1–11, Jan. 2012.

[17] C. Seibold, A. Hilsmann, and P. Eisert, "Style your face morph and improve your face morphing attack detector," in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Sep. 2019, pp. 1–6.

[18] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *J. Vis.*, vol. 16, no. 12, p. 326, 2016.

[19] A. Wolf, *ICAO: Portrait Quality (Reference Facial Images for MRTD), Version 1.0. Standad.* Montreal, QC, Canada: International Civil Aviation Organization, 2018.

[20] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.

[21] G. Borghi, E. Pancisi, M. Ferrara, and D. Maltoni, "A double Siamese framework for differential morphing attack detection," *Sensors*, vol. 21, no. 10, p. 3466, May 2021.

[22] C. Shu, X. Ding, and C. Fang, "Histogram of the oriented gradient for face recognition," *Tsinghua Sci. Technol.*, vol. 16, no. 2, pp. 216–224, Apr. 2011.

[23] S. Liao, X. Zhu, Z. Lei, L. Zhang, and S. Z. Li, "Learning multi-scale block local binary patterns for face recognition," in *Proc. Int. Conf. Biometrics*. Cham, Switzerland: Springer, 2007, pp. 828–837.

[24] J. Kannala and E. Rahtu, "BSIF: Binarized statistical image features," in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, Nov. 2012, pp. 1363–1366.

[25] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, Jul./Aug. 2008.

[26] U. Scherhag, C. Rathgeb, and C. Busch, "Towards detection of morphed face images in electronic travel documents," in *Proc. 13th IAPR Int. Workshop Document Anal. Syst. (DAS)*, Apr. 2018, pp. 187–192.

[27] U. Scherhag, C. Rathgeb, and C. Busch, "Morph deterction from single face image: A multi-algorithm fusion approach," in *Proc. 2nd Int. Conf. Biometric Eng. Appl. (ICBEA)*, 2018, pp. 6–12.

[28] L.-B. Zhang, F. Peng, and M. Long, "Face morphing detection using Fourier spectrum of sensor pattern noise," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–6.

[29] R. Raghavendra, K. B. Raja, S. Venkatesh, and C. Busch, "Transferable deep-CNN features for detecting digital and print-scanned morphed face images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1822–1830.

[30] S. Venkatesh, R. Ramachandra, K. Raja, and C. Busch, "Single image face morphing attack detection using ensemble of features," in *Proc. IEEE 23rd Int. Conf. Inf. Fusion (FUSION)*, Jul. 2020, pp. 1–6.

[31] S. Cai, L. Zhang, W. Zuo, and X. Feng, "A probabilistic collaborative representation based approach for pattern classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2950–2959.

[32] M. Ferrara, A. Franco, and D. Maltoni, "Face morphing detection in the presence of printing/scanning and heterogeneous image sources," *IET Biometrics*, vol. 10, no. 3, pp. 290–303, May 2021.

[33] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Jan. 2009.

[34] 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.

[35] M. Ferrara, A. Franco, and D. Maltoni, "Face demorphing," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 4, pp. 1008–1017, Apr. 2018.

[36] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: http://arxiv.org/abs/1411.1784

[37] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[38] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML Workshop Deep Learn. Audio, Speech, Lang. Process.*, 2013, p. 3.

[39] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[40] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. ICML*, 2010, pp. 807–814.

[41] C. Li and M. Wand, "Precomputed real-time texture synthesis with Markovian generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 702–716.

[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015, pp. 1–15.

[43] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in Neural Information Processing Systems*, vol. 27, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2014, pp. 2366–2374.

[44] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1. Jun. 2005, pp. 947–954.

[45] S. Pini, G. Borghi, and R. Vezzani, "Learn to see by events: Color frame synthesis from event and RGB cameras," in *Proc. 15th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, 2020, pp. 37–47.

[46] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," 2016, *arXiv:1606.03498*. [Online]. Available: http://arxiv.org/abs/1606.03498

[47] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, ''Image quality assessment: From error visibility to structural similarity,'' *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[48] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, ''The unreasonable effectiveness of deep features as a perceptual metric,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[49] K. Simonyan and A. Zisserman, ''Very deep convolutional networks for large-scale image recognition,'' 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, ''ImageNet: A large-scale hierarchical image database,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[51] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, ''ArcFace: Additive angular margin loss for deep face recognition,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4685–4694.

[52] *FVC Ongoing*. Accessed: Jul. 6, 2021. [Online]. Available: https://biolab.csr.unibo.it/fvcongoing/UI/Form/Home.aspx

[53] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

GUIDO BORGHI received the M.Sc. degree in computer engineering and the Ph.D. degree in information and communication technologies from the University of Modena and Reggio Emilia, Italy, in 2015 and 2019, respectively. He is currently an Assistant Professor with the Department of Computer Science and Engineering, University of Bologna, Cesena Campus. His research interests include computer vision and deep learning techniques applied to intensity and depth images for face analysis, biometrics, driver monitoring, and human–computer interaction.

ANNALISA FRANCO received the Ph.D. degree in electronics, computer science, and telecommunications engineering from DEIS, University of Bologna, Italy, in 2004, with a focus on multidimensional indexing structures and their application in pattern recognition. She is currently an Associate Professor at the Department of Computer Science and Engineering, University of Bologna. She is a member of the Biometric System Laboratory (computer science), Cesena. She has authored several scientific papers and served as a referee for a number of international journals and conferences. Her research interests include pattern recognition, biometric systems, image databases, and multidimensional data structures. Her recent research activity is mainly focused on face recognition in the context of electronic identity documents.

GABRIELE GRAFFIETI received the degree *(cum laude)* from the University of Bologna, in 2018, where he is currently pursuing the Ph.D. degree in data science and computation with the Computer Science and Engineering Department, under the supervision of Prof. D. Maltoni. His main research interests include artificial intelligence and machine learning, in particular generative models (generative adversarial networks) and continual/lifelong deep learning.

DAVIDE MALTONI (Senior Member, IEEE) is currently a Full Professor at the Department of Computer Science and Engineering (DISI), University of Bologna. He is also the Co-Director of the Biometric Systems Laboratory (BioLab), which is internationally known for its research and publications in the field. Several original techniques have been proposed by BioLab Team for fingerprint feature extraction, matching, and classification, for hand shape verification, face location, and performance evaluation of biometric systems. He is the coauthor of the *Handbook of Fingerprint Recognition* (Springer, 2009) and holds three patents on fingerprint recognition. His research interests include pattern recognition, computer vision, machine learning, and computational neuroscience. He was elected as the International Association for Pattern Recognition (IAPR) Fellow, in 2010.