De-Occlusion Face Model based on Deep Occlusor Segmentation and Deep Inpainting Models

Miguel A. Gutierrez-Velazquez, Mario I. Chacon-Murguia, *Senior, IEEE*, and Juan A. Ramirez-Quintana

***Abstract*— Image inpainting is a computer vision task that reconstructs missing image regions. Given its potential for various applications, it is an area of great interest. Despite advances in this field thanks to deep models such as autoencoders and generative adversarial networks, fundamental challenges persist, such as the causal interpretation of information loss and the risk of overfitting and lack of diversity in the features obtained with autoencoders. In this context, this article presents an innovative deep network model to solve occluded face inpainting. The model focuses on attributing the loss of information to the occlusion. The proposed model consists of two deep models: one for segmenting the object occluding the face, called *SOCLNET*, and another for reconstructing the face, *IFACENET*. *SOCLNET* is an improvement of the DeepLabv3 network by adding self-attention mechanisms. *IFACENET* is based on an autoencoder with an ensemble learning approach in the encoder to improve the diversity of the extracted features. *SOCLNET* was evaluated to demonstrate that the segmentation of occluding objects works adequately, even on out-of-distribution images. Its performance metrics were Pixel Accuracy = 0.93 and IoU = 0.788. The *IFACENET* model was compared against other state-of-the-art models using the Celeb-HQ database. The quantitative results of *IFACENET* show an average performance of SSIM = 0.95, PSNR = 26.813, and L1 = 0.261 with different mask values, being competitive with the state of the art. Additionally, qualitative results of *IFACENET* are shown to demonstrate the visual outcomes of face inpainting. Based on those results, it can be concluded that the proposed model effectively solves the reconstruction of occluded faces, opening new perspectives in the research of image reconstruction.**

Link to graphical and video abstracts, and to code: https://latamt.ieeer9.org/index.php/transactions/article/view/9612

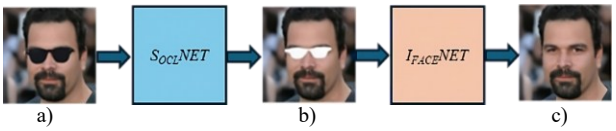
***Index Terms*—Deep learning, face reconstruction, image inpainting, occluded objects segmentation.**

I. Introduction

IMAGE inpainting is a field of computer vision primarily aimed at producing a visually plausible structure in missing regions of images or restoring damaged pixels in an image [1], [2]. Significant advances have been achieved in image inpainting due to the progress of deep learning, especially with autoencoders [3], [4] and generative networks [5], [6]. Models are typically based on autoencoders [7], [8] or in combination with generative networks [9]-[12]; attention mechanisms [13] and transformers have also been recently integrated into image inpainting [14], [15].

Some of these existing models have shown competent results for reconstructing images with regular or irregular masks and of different sizes [16]-[19]. However, two critical issues remain unresolved. First, conventional approaches rely on user-supplied or artificially generated masks, overlooking the genuine cause of missing information—actual occluding objects (e.g., glasses, hats) that hide key facial regions. This reliance on manual or random masks means the algorithm never automatically attributes missing areas to specific occluders. On the other hand, approaches that use an autoencoder as a generator [11], [19]-[22] extract features from the images using a single model, risking overfitting and limited feature diversity. Such an encoder can capture only a narrow subset of features, leading to bias or suboptimal reconstructions when the missing region is extensive or suboptimal manifold representation in the latent space. Therefore, this work proposes a new deep model to address these issues. The problems mentioned earlier, and the proposed solutions are explained in more detail below, and Fig. 1 shows the proposed general model, which first consists of a model called *SOCLNET* that segments the occluding object and then a model called *IFACENET* that reconstructs the missing part of the face.

Problem 1. Cause of information loss. Image inpainting models reconstruct an image I to which a mask m is applied, causing I to lose information and obtain an image $I' \approx I$. In the state-of-the-art literature, it is assumed that a m was applied to I, or in various applications, a causal sense is given to the lack of information, but not automatically. In this work, we delve deeper into this topic and consider that the loss of information is due to occlusion. To justify this approach, let's formalize the problem. Assume that there is a complete object of interest O in I. Now consider that object O is occluded by some occluding object O_{ocl}. In this case, there is a loss of information, so the image can no longer be I, but it also cannot be I_m, because O_{ocl} is not necessarily a m; since $O_{ocl} \in \{x \mid Y(x)\}$, that is, an occluding object is an object x (which can be glasses, hands, cameras, face masks, etc.) that satisfies the condition $Y(x)$. Thus, the image that lacks information due to O_{ocl} is defined as I_{inc}. Therefore, the object of interest O is incomplete and is termed O_{inc}. The problem is obtaining an image I' from I_{inc} and not just from I_m.



a) b) c)

Fig. 1. Proposed general model. a) Occluded face, b) occlusor segmentation, c) face reconstruction.

The associate editor coordinating the review of this manuscript and approving it for publication was Samuel Ortega (*Corresponding author: Miguel Gutierrez*).

Miguel Gutierrez, M. Chacon, and J. Ramirez are with the Visual Perception Systems Lab, Tecnologico Nacional de Mexico/I.T Chihuahua, Chihuahua, Mexico (e-mails: m19061419@chihuahua.tecnm.mx, mario.cm@chihuahua.tecnm.mx, juan.rq@chihuahua.tecnm.mx).

To solve this problem, a new model is proposed to reconstruct the object of interest, *O*, which is a face in this work. This proposed model involves: a) $S_{OCL}NET$ designed to segment occluding objects such as $S_{OCL}NET$: $I_{inc} \rightarrow I_m$; and b) a face reconstruction model such that $I_{FACE}NET$: $I_m \rightarrow I'$.

Problem 2. Risk of overfitting, bias, or lack of diversity in the encoder stage. Let *E* be the encoder that extracts features, *F*, from I_m, through convolutional layers, such that $E(I_m) = F$, where $F \in \mathbb{R}^{cxwxh}$. *F* may have features with overfitting, bias, or lack of diversity. Therefore, employing an ensemble learning approach in the encoder is advisable to deal with those issues. In this work, the stacking method was selected. This method consists of combining features of two or more models, indicted as models of level 0 and a model of level 1, which learns to combine the features better [23].

Based on the problems mentioned 1) and 2) and their respective solutions, the contributions of this paper are:

- The work proposed a new model to complete faces occluded by different objects, thus extending the applications of image inpainting.
- A new model to segment occluding objects $S_{OCL}NET$ is designed by improving the DeepLabv3 model with a self-attention mechanism.
- An autoencoder-based face reconstruction model, $I_{FACE}NET$, is designed, employing ensemble learning in the encoder to acquire more general and representative features.
- Evaluation of $I_{FACE}NET$ on the Celeb-HQ database demonstrates that $I_{FACE}NET$ is competitive quantitatively and qualitatively concerning state-of-the-art models.
- In-depth analysis of the benefits of the ensemble learning approach in the inpainting model $I_{FACE}NET$.

The rest of the paper is organized as follows: Section II provides a literature review. Section III explains the proposed model. Section IV presents the experiments and results. Finally, Section V discusses the conclusions.

II. Related Works

Image inpainting methods can be divided into two approaches: traditional methods and deep learning-based methods [24]. This section describes deep learning-based methods and reviews some techniques for reconstructing occluded faces. Interested readers are referred to the following works for traditional methods [25].

A. Image Reconstruction with Deep Learning

Significant improvement in image reconstruction models has been achieved through the advancement of deep learning [1] and adversarial training [26], which enables the extraction of more meaningful semantic information [27]. Deep learning-based methods are divided into three types: single-stage, two-stage, and progressive [28]. In this section, we will mention the single-stage methods.

The single-stage methods learn to map an incomplete image to a complete one. They typically consist of a single generator along with their respective cost functions. A seminal work was conducted by D. Pathak *et al*. [7]. In this work, the reconstruction model is an autoencoder with two cost functions: an L_2, and an adversarial cost function, which depends on a discriminator, aiming to reconstruct an image at the pixel level like the original image. Building upon this previous work in [8], they employ an autoencoder and two global and local discriminators to capture finer details. In [9] the authors use ResNet as the generator. To capture fine details, they utilize an image patch-based discriminator. Chen *et al*. [21] also utilize an autoencoder and a discriminator, but they include residual connections. Yan *et al*. [29] add a shift connection to the generator with residual connections to enhance the reconstruction. In addition to residual connections, in [16], they implement regular convolutions and DenseNet-based convolutions in the generator along with attention mechanisms. Liu *et al*. [30] propose the use of partial convolutions for a more realistic reconstruction. In [31] they enhance partial convolutions by adding bidirectional attention maps. Zeng *et al*. [19] introduce AOT-Blocks, inspired by [32] for extracting contextual information. In this scheme, convolutions with different dilation factors are concatenated. A similar approach is demonstrated in [33], where features from convolutions with different dilation factors are combined with an attention layer. In [34], they merge low-level and high-level semantic features for contextual information extraction, allowing the model to learn how to reconstruct missing parts. The use of attention mechanisms for contextual information extraction has also been proposed. The first work to add attention mechanisms for image reconstruction was [10]. This work inspired several subsequent works, such as that of Qin et al. [35], where they propose a multi-scale attention mechanism. In [36], a network is presented to learn semantic information, along with a residual connection attention mechanism for image reconstruction.

Despite the previous approaches achieving good results, there is still more research, such as using ensemble learning and methodologies to give causal meaning to occluded information automatically.

B. Occluded Face Inpainting

Significant progress has been made in the task of occluded face inpainting. This section presents a variety of innovative methods employed for reconstructing occluded faces.

Ge *et al*. [37] present a scheme composed of three parts: a generator consisting of an autoencoder and a network for extracting semantic information, a discriminator, and a recognition network that adds a term to the cost function; however, the occlusion of faces is limited to a binary rectangle. This limitation is also present in [38], [39].

Yang *et al*. [40] propose a model based on an autoencoder for reconstructing occluded faces, where gray rectangles in different face positions represent occlusions. In [41], a methodology based on a GAN trained with images of non-occluded faces and a network to detect the occluding object. However, the only detected occluding objects are lenses. Jiang et al. [42] employ a model based on CGAN (Conditional Generative Adversarial) to reconstruct parts of the face occluded with different objects. A methodology to eliminate occlusion with a model and another model to reconstruct the occluded region is described in [43]. Two GANs are proposed in [44]; one finds the occluded region to be completed in the

face and the other to reconstruct that region. D. Kim and U. Park [45] present a model with an autoencoder with residual connection. The input to the model is an occluded face and a binary mask generated with a Laplacian operator applied over the original image. In [50], the original face image is divided into patches, and each patch is automatically classified as an occluded or non-occluded region. Then, the occluded regions are reconstructed. Li *et al*. [47] show a methodology with Long-Short Time Memory to find the occluded region from critical points to locate a human face. Xu *et al*. [48] describe a model to reconstruct occluded faces. They generate different reconstruction techniques from an attention mechanism termed Parallel Visual Attention. In [49] the faces occluded by different objects are reconstructed by combining segmentation and an autoencoder-type generator with residual connections to reconstruct the occluded face.

Considering the previously analyzed works, face reconstruction presents several limitations. The occluding object is considered only a rectangle with specific pixel values, or there are only a few types of occluding objects. The paradigms that first segment the occluding object and then reconstruct the face are promissory. However, these approaches do not warrant that the segmented objects are occluding objects.

III. Description of the Proposed Models $S_{OCL}NET$ and $I_{FACE}NET$

The new model proposed in this paper consists of two models shown in Fig. 1 $S_{OCL}NET$ and $I_{FACE}NET$. These models are designed with specific and complementary purposes. $S_{OCL}NET$ aims to segment the object that occludes the face, while $I_{FACE}NET$ is focused on reconstructing the occluded face region in the image. This section describes the architecture of both models and the implemented cost functions.

A. Segmentation Model, SOCLNET

The objective of $S_{OCL}NET$ is to segment occluding objects. Its architecture is based on the state-of-the-art model DeepLabV3 [32]. This model was chosen for its ability to extract contextual information, which is essential for distinguishing between occluding and non-occluding objects. DeepLabV3 involves a backbone network, which in this work is the ResNet50 network [50], and the Atrous Spatial Pyramid Pooling (ASPP) module to capture contextual information. Self-attention mechanisms have been added to DeepLabv3, as shown in Fig. 2, to improve the contextual information. The self-attention mechanism is given by

$$F_{attention} = \text{Attention}\left(F_{input}\right) = \lambda\sigma\left(QK^{T}\right)V + F_{input}, \quad (1)$$

where λ is a learning parameter, σ is the sigmoid activation function, Q is query, K is key, V is value and F_{input} is the input feature map from which the most important features will be found. To complete the description of this model, the cost function used in $S_{OCL}NET$ is the binary cross-entropy (BCE) between the segmentation result and the ground truth provided in the database.

B. Face Reconstruction Model, $I_{FACE}NET$

The general structure of $I_{FACE}NET$ is illustrated in Fig. 3. $I_{FACE}NET$ includes several stages: a) an autoencoder, which includes an encoder employing ensemble learning, b) a contextual information extraction block based on [19] and [32], c) a decoder, and d) a discriminator. The encoder's input consists of the original image, I, and the mask, m, and the following operation is performed with these images.

$$I_m = I \oplus (1-m) + m \quad (2)$$

where $\oplus$ is the Hadamard product. The encoder consists of two level 0 models and one level 1 model, formally

$$E(I_m) = M_1^1\left(M_1^0(I_m), M_2^0(I_m)\right) \quad (3)$$

Since the models M_1^0 and M_2^0 extract a feature map F_1^0 and F_2^0 respectively, then

$$E(I_m) = M_1^1\left(F_1^0, F_2^0\right) = F_{N1} \quad (4)$$

where the feature map generated by the autoencoder leads to a better generalization, representation, and diversity of features. The encoder is used to increase the diversity of features extracted by combining a very deep network, TraResNet (M_1^0), based on ResNet50 [50], with a shallow network called Shallower (M_2^0) with an attention-based model (M_1^1). This is crucial for capturing different features from input images. The TraResNet network is the ResNet50 network from which the classification layers are removed, and two transposed convolutional layers are added to increase the spatial dimension. ResNet50 is used due to its low condition number [52], which indicates a more stable optimization process; it also alleviates singularity issues [53] and has a smoother error surface [54], leading to better generalization. The ResNet50 network is already pre-trained, so only its last two layers are fine-tuning to adapt to the new task. The Shallower network consists of 4 layers of depthwise separable convolutions, considerably less than ResNet50. The depthwise separable convolutions decrease training time by performing operations more efficiently and reducing the number of parameters [55]. The ReLU activation function follows each depthwise separable convolution. Equation (5) defines the mathematical operation performed by the level 0 models, TraResNet and Shallower.

$$F_{N0} = \text{TraResNet}(I_m) \otimes \text{Shallower}(I_m) \quad (5)$$

where F_{N0} is the feature map obtained by concatenating the networks TraResNet and Shallower outputs. The level 1 model consists of an attention mechanism like (1), such that

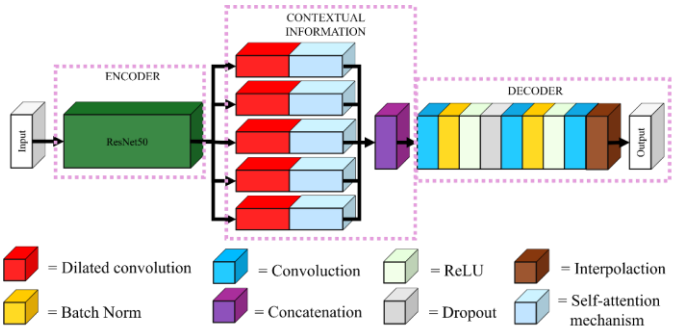


Fig. 2. $S_{OCL}NET$ architecture. The ResNet50 feature extractor acts as an encoder, its output features pass through the contextual information extraction stage, and finally, an encoder gets the input's dimension size back.

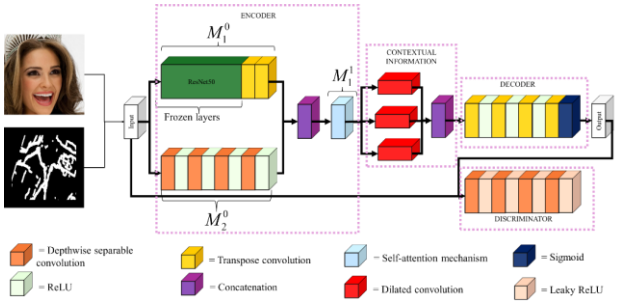


Fig. 3. $I_{FACE}NET$ architecture. Two level 0 models M_1^0 and M_2^0 extract different features that are concatenated. The level 1 model, M_1^1 combining their features. The encoder output goes into a contextual information extraction stage, followed by the encoder. A discriminator adds a term in the loss function during training but is absent during inferences.

$$F_{N1} = \text{Attention}\left(F_{N0}\right) \tag{6}$$

where F_{N1} is the feature map of the level 1 model. Therefore, the encoder, E, of the model $I_{FACE}NET$ performs the operation $E(I_m) = F_{N1}$. Thus, the feature map F_{N1} is computed by learning to combine the features of the Shallower and ResNet50 models. On the other hand, the contextual extraction block consists of three depthwise separable convolutions with dilation factors of $r = 1$, 2, and 4, which are concatenated.

Since the input size decreases, it does not correspond to the dimension of the original image $I_m \in \mathbb{R}^{3x256x256}$; it is necessary to include a decoder made up of transposed convolutions to recover the dimension of the original image. Finally, the discriminator, D: $D(I) = F_{patch}$, has five convolutions. $F_{patch} \in \mathbb{R}^{30x30}$ is the output feature map of the discriminator. Each element of F_{patch} represents a specific region of the image. The discriminator contributes to the cost function with a term during training. The cost function of the $I_{FACE}NET$ must ensure that its minimization produces adequate pixel-by-pixel reconstruction and visual fidelity. Therefore, the following cost functions are chosen: L_{rec}, L_{per}, L_{sty} and L_{adv}. The L_{rec} cost function ensures a correct pixel-by-pixel reconstruction in the reconstructed image.

$$L_{loc} = \left\| I' - I \right\|_1 \tag{7}$$

The perceptual cost function, L_{per}, and the style cost function, L_{sty}, aim to reduce the error between the generated and real images at the feature level rather than at the pixel level [56]. L_{per} is defined as

$$L_{per} = \frac{\left\| \phi_j\left(I'\right) - \phi_j\left(I\right) \right\|}{N_j}, \tag{8}$$

where ϕ_j is the activation map of the j-th layer of the pretrained network VGG19 [57]. VGG19 was used because it is broadly employed to generate activation maps to compute L_{per} image [19]. $\phi_j \in \mathbb{R}^{cxhxw}$, and $N_j = cxhxw$. N_j is the number of elements of the activation map in the layer j.

L_{sty} is the distance between the Gram matrices, representing the correlation between various features. The correlation between multiple features defines the style.

$$L_{sty} = \frac{\left\| \phi_j\left(I'\right)^T \phi_j\left(I'\right) - \phi_j\left(I\right)^T \phi_j\left(I\right) \right\|}{N_j} \tag{9}$$

Finally, L_{adv}, corresponds to the binary cross entropy BCE of the feature maps and F_{patch}, obtained by the discriminator from the input images I' and I

$$L_{adv} = \text{BCE}\left(F_{patch}^{'}, F_{patch}\right) \tag{10}$$

The total cost function is a weighted sum of the aforementioned cost functions.

$$L_R = \lambda_{rec} L_{rec} + \lambda_{sty} L_{sty} + \lambda_{per} L_{per} + \lambda_{adv} L_{adv} \tag{11}$$

To select the λ_i hyperparameters we employed Optuna [58] an automatic hyperparameter optimization framework to optimize the hyperparameters associated with the loss function. The objective function, defined within the Optuna study, takes as input a set of sampled hyperparameters (i.e., the loss weights λi), trains the model using these hyperparameters, and returns the total validation loss as the scalar value to be minimized. We first performed a random search during 40 trials to explore the hyperparameters space, followed by 10 trials using the Tree-structured Parzen Estimator (TPE) [59]. The search space for each loss weight was set between 1.0 and 10.0 with an increment of 0.1. This approach leads to the following loss weights: λrec = 1.7, λper = 4, λsty = 3.6, and λadv = 4.7.

IV. Experiments and Results

This section presents the experiments conducted using the $S_{OCL}NET$ and $I_{FACE}NET$ models. The explanation begins with a mention of the architectures, hyperparameters, and optimizers for each model used in the experiments. Then, the databases used for training and the metrics selected to measure the performance of the models are described. Then, the results obtained for the two models, $S_{OCL}NET$ and $I_{FACE}NET$, are presented. Finally, the reconstruction of occluded faces is demonstrated visually.

A. Experiments and Results of $S_{OCL}NET$

Two experiments were conducted with the $S_{OCL}NET$ model: one with the DeepLabv3 model without attention mechanism, called $S_{context}$, and another with the with attention mechanism, called S_{conatt}. Both experiments were carried out for 20 epochs, with a learning rate of 0.001, using the SGD optimizer and a batch size of 2. The resolution of the training images was 640x640.

a) Database

The database used for $S_{OCL}NET$ is based on the RealOcc database [60]. The choice of RealOcc is justified for two key reasons: a) it represents real-world conditions, and b) it has a wide diversity of occlusions. RealOcc is a database for segmenting occluded faces, thus, the occluding object mask is manually added to the ground truths. This process generates a new database called RealOccOcl. Adding the occluding class provides important contextual information for a better understanding of the entire image to help the model learn the condition $Y(x)$. The training set of RealOccOcl consists of 214 images, and the validation set contains 54 images, each with its respective ground truths.

b) Metrics

The evaluation of a segmentation model should involve measuring both pixel classification accuracy and correct localization. Therefore, the accuracy and intersection over union (IoU) metrics were computed.

c) Results

To our knowledge, no other work has utilized the RealOccOcl database. Hence, we trained the following state-of-the-art segmentation models with RealOccOcl dataset: FCN, UNet++, AttnUNet, PSPNet, FPN, PANet, and ViT. The quantitative results are shown in TABLE I. The model S_{conatt} has the highest IoU value. In Fig. 5 it can be seen that S_{conatt} is capable of accurately segmenting various types of occluders.

However, such capability must remain consistent across another database. Various inferences were made on out-of-distribution (OOD) images to assess the generalization quality and determine whether the solution is a shortcut or an OOD solution [61]. These inferences involve passing OOD images through the S_{conatt} model with objects x that satisfy $Y(x)$ and with the same object x that does not satisfy $Y(x)$. Observing some inferences illustrated in Fig. 4 we can conclude that the model can segment faces and occluding objects in OOD images.

TABLE I
COMPARISON OF THE OCCLUDED SEGMENTATION TASK RESULTS

	Model								
Metrics	FCN	Unet++	AttnUnet	PSPNet	FPN	PANet	ViT	$S_{context}$	S_{conatt}
PAcc	0.9	0.85	0.9	0.88	0.94	0.87	0.96	0.91	0.93
IoU	0.58	0.65	0.76	0.5	0.74	0.48	0.72	0.76	0.78

a)

b)

Fig. 5. Some results of S_{OCLNET} on validation images from RealOccOcl. The face is shown in white, and the occluding object is shown in blue. a) Original image, b) segmentation result.

However, due to the intrinsic limitation of validating the S_{conatt} model on all possible images, adopting additional methods to justify and understand how the model makes its predictions and thus establish a framework of confidence and transparency becomes imperative. This was achieved by implementing an XAI (Explainable Artificial Intelligence) technique [62], specifically Grad-CAM (Gradient-weighted Class Activation Mapping) [63]. Grad-CAM is used to show the areas the model considers relevant for segmentation and validate its performance qualitatively. This is achieved through the following equations:

a)

b)

Fig. 4. Examples of inferences from S_{OCLNET} in OOD images. Faces are blanked out, and occluding objects are highlighted in blue. a) Input OOD images with objects, b) segmentation result.

$$\alpha_k^c = \frac{1}{z}\sum_i\sum_j \frac{\partial S_{conatt}^c}{\partial f_{kij}} \tag{12}$$

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c f_k\right) \tag{13}$$

The equation (12) calculates the importance weights α_k^c for class c in each feature map k. The mean gradient score of the class c determines these weights, S_{conatt}^c, concerning each pixel (i,j) of the feature map, f_{kij}. The gradient reflects how modifying a pixel in the feature map would affect the class score, which provides the relevance of each location in the feature map for classification. On the other hand, equation (13) provides an activation map for class c, $L_{\text{Grad-CAM}}^c$, which is obtained by applying the ReLU function to the weighted sum of the feature maps f_k using the importance weights α_k^c. This ensures that only the pixels positively influencing the class score S_{conatt}^c are considered, eliminating those with a negative effect. The result can be visualized with a heatmap overlaid on the original image to reveal the regions the model considers decisive for predicting the class.

The heatmaps are shown in Fig. 6 highlight the regions the S_{conatt} model considers most important for classifying the different classes: background, face, and occluder. To predict the background class, the model prioritizes the peripheral regions around the face, which would be expected. In contrast, the

prediction of the face class shows adaptable attention to various facial attributes, such as the nose and the areas adjacent to the hair or eyes. Thus, the model has other reference points for correct facial segmentation if an important landmark is occluded.

Regarding the prediction of the occluder class, the model pays intense attention to the facial region. This focus serves a dual purpose: determining the proximity of an object to the face to ascertain if it's an occluder and distinguishing that object. The application of Grad-CAM enhances the model's segmentation understanding, extending the evaluation beyond quantitative metrics and qualitative demonstrations.

Once the generalization of the model has been validated, the final step is to ensure that $S_{OCL}NET$: $I_{inc} \rightarrow I_m$, as it is required that in the output image, only the occluded region is segmented. This way, the mask in I_m corresponds to the segmented region with an occluding object. To achieve this, it is sufficient to use only the output of the occluder class on the original image I to obtain I_m.

B. Experiments and Results of $I_{FACE}NET$

This section describes the experiments conducted with $I_{FACE}NET$ and its quantitative and qualitative results. The model was trained for 20 epochs using the Adam optimizer, a learning rate 0.0001, and a batch size of 8. The input images are 3x256x256.

a) Dataset

Due to its extensive usage in state-of-the-art image reconstruction tasks, the Celeb-HQ [64] database was chosen for these experiments. Celeb-HQ includes 30000 images of human faces with dimensions of 1024x1024. All images are resized to 256x256. The shared masks from [30] are used as masks.

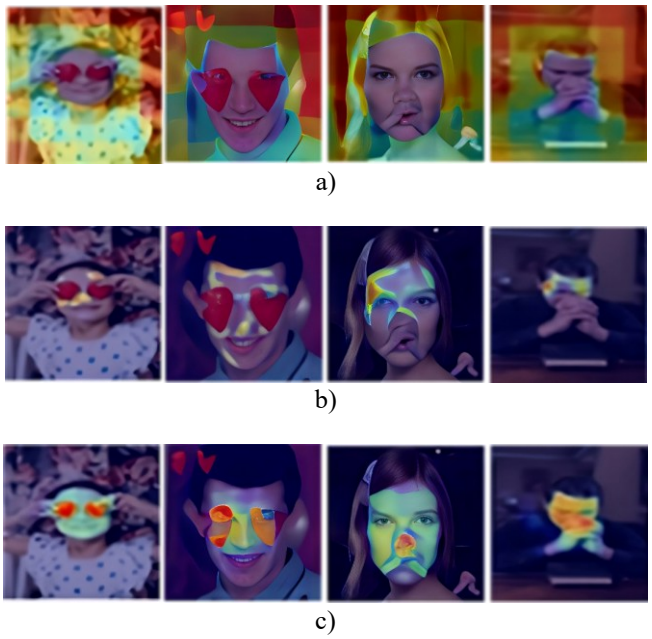


Fig. 6. Heatmaps obtained with Grad-CAM. a) Prediction: background, b) prediction: face, c) prediction: occluder.

Like [30], the database is divided into 27000 images for training and 3000 for validation, each with its respective ground truths.

b) Metrics

The evaluation of the reconstructed images is divided into two assessments: one is based on human subjectivity, which is the most effective and consists of qualitative assessment [19], and quantitative assessments. For this reason, several inferences are conducted to demonstrate the visual performance of the proposed model $I_{FACE}NET$. Quantitative evaluations are obtained with metrics that should represent both accurate pixel-level reconstruction and structural similarity. Therefore, the following quantitative metrics are used:

- $\ell 1$. It is a metric that compares the difference between the value of the pixels in the original image and the reconstructed image. This metric has been widely used in other works [7], [11], [19]. A low value leads to a higher pixel-level resemblance.
- Peak Signal-to-Noise Ratio (PSNR): It is typically used to measure the quality of a reconstructed signal (reconstructed image) compared to its original reference signal (original image). A high value indicates a more significant similarity between the compared images.
- Structural Similarity Index (SSIM): This metric evaluates how similar the original and reconstructed images are based on three factors: illumination, contrast, and structure. The closer it is to 1, the greater the resemblance between the two images.
- Fréchet Inception Distance (FID): This metric measures the distance between the feature distributions of the original and reconstructed images, extracted using a pre-trained Inception network. Lower FID values indicate that the reconstructed images are statistically closer to the real images in terms of high-level features and, thus, perceptually more realistic.
- Learned Perceptual Image Patch Similarity (LPIPS): Evaluates perceptual similarity by comparing deep features from a pretrained AlexNet network. A lower LPIPS score corresponds to a higher perceptual similarity.

c) Results

To evaluate the $I_{FACE}NET$ model, a quantitative comparison was carried out using the Celeb-HQ database [64] against the following models: CA [10], PConv [30], GatedConv [65], HiFill [66], AOT [19], and Local-Global Refinement [11]. Performance metrics were obtained using the validation set with masks of different sizes; these bins are 10%-20%, 20%-30%, 30%-40%, 40%-50%, and 50%-60%. These percentages indicate the percentage of pixels corresponding to the binary mask. The quantitative results are shown in TABLE II, The comparative analysis from TABLE II reveals that our model is competitive compared to other state-of-the-art models. Specifically, $I_{FACE}NET$ is the only model ranking among the top two in all metrics. Also, in [14] a transformer-based inpainting model is present, and its results are within our model´s range: SSIM = 0.941, LPIPS = 0.079, and FID = 4.46. On the other hand, the qualitative results are shown in Fig. 7. The reconstruction performed by $I_{FACE}NET$ on various faces with different characteristics demonstrates that it is visually plausible under different mask sizes by using current information and prior knowledge.

C. Ablation Studies

To thoroughly assess the importance of each component in our ensemble encoder, we present a three-part ablation study. First, we compare the proposed ensemble model, $I_{FACE}NET$,

TABLE II
COMPARISON OF THE RESULTS OF I_{FACE}NET WITH OTHER MODELS USING THE CELEB-HQ DATABASE. THE TWO BEST RESULTS REGARDING THE PERCENTAGE OF OCCLUSION ARE HIGHLIGHTED. THE METRICS ↑ INDICATES THAT A HIGHER VALUE IS BETTER AND ↓ INDICATES THAT A LOWER VALUE IS BETTER

		Inpainting models						
Metric	Mask	CA [10]	PConv [30]	Gated Conv [65]	HiFill [66]	AOT [19]	Local-Global [11]	*I_{FACE}NET*
L_1 (10^{-2}) ↓	1-10%	0.89	0.68	0.66	0.67	0.55	**0.46**	**0.34**
	10-20%	2.07	1.28	1.55	1.81	**1.19**	1.28	**0.89**
	20-30%	3.54	2.42	2.73	3.31	**2.11**	2.38	**1.61**
	30-40%	5.19	3.43	4.08	5.02	**3.2**	3.72	**2.42**
	40-50%	7.07	4.62	5.68	7.12	**4.51**	5.27	**3.39**
	50-60%	10.11	7.74	8.09	10.47	**7.07**	8.38	**5.28**
PSNR↑	1-10%	31.07	34.04	32.95	30.97	34.79	**40.04**	**34.8**
	10-20%	25.81	28.75	27.05	25.36	29.49	**33.99**	**30.1**
	20-30%	22.93	25.59	23.81	22.35	26.03	**30.54**	**27.1**
	30-40%	20.98	23.4	21.55	20.21	23.58	**27.99**	**25.03**
	40-50%	19.23	21.56	19.75	18.35	21.65	**26.01**	**23.94**
	50-60%	17.1	18.75	16.94	16.07	19.01	**23.12**	**20.62**
SSIM↑	1-10%	0.961	0.971	0.97	0.963	0.976	**0.995**	**0.9957**
	10-20%	0.906	0.928	0.921	0.905	0.94	**0.98**	**0.9911**
	20-30%	0.8444	0.875	0.86	0.835	0.89	**0.96**	**0.9851**
	30-40%	0.783	0.82	0.79	0.762	0.835	**0.94**	**0.9755**
	40-50%	0.72	0.762	0.727	0.68	0.773	**0.917**	**0.9656**
	50-60%	0.648	0.677	0.626	0.588	0.682	**0.849**	**0.9407**
LPIPS↓	1-10%	-	-	0.012	-	-	**0.006**	**0.001**
	10-20%	-	-	0.034	-	-	**0.017**	**0.02**
	20-30%	-	-	0.061	-	-	**0.031**	**0.022**
	30-40%	-	-	0.091	-	-	**0.048**	**0.045**
	40-50%	-	-	0.125	-	-	**0.069**	**0.07**
	50-60%	-	-	0.181	-	-	**0.108**	**0.2**
FID↓	1-10%	1.3	0.36	0.21	0.53	**0.20**	0.39	**0.35**
	10-20%	6.33	1.85	0.85	2.52	**0.61**	1.06	**0.8**
	20-30%	17.36	5.83	2.4	7.60	**1.57**	2.08	**2.03**
	30-40%	34.26	12.96	5.33	17.18	3.38	**3.16**	**3.21**
	40-50%	56.89	24.63	10.66	36.23	6.89	**4.61**	**4.46**
	50-60%	82.67	47.09	32.9	72.03	20.20	**7.07**	**6.07**

Fig. 7. a) Original image, *I*. b) Image with a binary mask, I_m. c) Image reconstructed by *I_{FACE}NET*.

against three ablation variants, each removing a specific sub-encoder or module. Next, we analyze their latent feature manifolds using Uniform Manifold Approximation and Projection (UMAP) [67] to gauge how each model organizes occluded inputs. Finally, we examine representational similarity via Centered Kernel Alignment (CKA) [68] to verify that each encoder element learns non-redundant features.

a) Performance Comparison of Ablation Variants

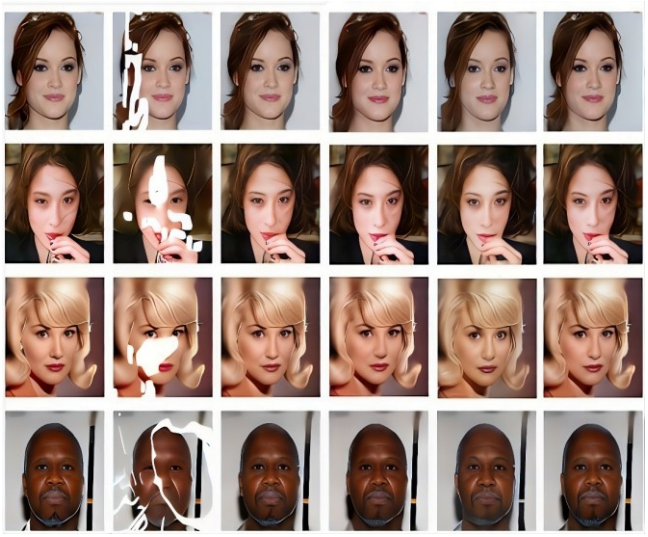
We trained and evaluated four models: *I_{FACE}NET*, and the three ablated versions, each omitting a distinct level 0 model. In TABLE III, we share the quantitative results. It is shown that *I_{FACE}NET* outperforms the ablated counterparts in all metrics. Qualitative analysis (Fig. 8) confirms that the ensemble yields more coherent and natural restorations, indicating robustness to missing pixel regions.

b) Feature Manifold Analysis Via UMAP

UMAP is a dimensionality-reduction technique that projects high-dimensional data into a lower-dimensional space while attempting to preserve local and global structure. Commonly used for clustering or class separation in classification tasks, UMAP can also provide insights into how a model organizes its learned features in a continuous manifold. In our inpainting scenario, we apply UMAP to the encoder's feature outputs across all different occlusion bins. We focus on derived metrics (silhouette score, Davies-Bouldin Index, PCA Variance) to assess the shape and stability of each model's latent space. We share the results in TABLE IV. *I_{FACE}NET* yields a less spread manifold by UMAP criteria, signaled by silhouette values and higher Davies-Bouldin indices. However, these metrics remain stable across bins, hence its low standard deviation value, implying the model *I_{FACE}NET* encodes occluded inputs in a cohesive latent distribution rather than fragmenting them. This stability aligns with the better reconstruction metrics, showing that a tighter manifold (in this inpainting context) correlates with stronger resilience under occlusion. On the other hand, ablation models produce more scattered or multiple sub-clusters. That implies their latent space fluctuates across bins, suggesting less consistent handling of occluded faces. Hence, spread in this scenario does not indicate improved generalization but rather an inability to unify partial inputs into a robust shared representation.

TABLE III
METRICS OF ABLATION STUDIES. ABLATIONV1 = MODEL WITHOUT. ABLATIONV2 = MODEL WITHOUT. ABLATIONV3 = MODEL WITHOUT. THE RESULTS ARE THE MEAN ACROSS ALL BINS OF OCCLUSION. METRICS ↑ INDICATES A HIGHER VALUE IS BETTER AND ↓ INDICATES A LOWER VALUE IS BETTER

Models	Metrics				
	$L_1(10^{-2})\downarrow$	PSNR↑	SSIM↑	LPIPS↓	FID↓
I_{FACE}NET	**2.3216**	**26.93**	**0.9756**	**0.0596**	**2.82**
AblationV1	2.51	26.23	0.95	0.2350	5.84
AblationV2	2.77	25.166	0.92	0.2391	7.46
AblationV3	2.36	26.71	0.972	0.2338	8.687



a) b) c) d) e) f)

Fig. 8. Comparison between I_{FACE}NET and ablated models. a) Original image, b) masked image, c) I_{FACE}NET result, d) AblationV1 result, e) AblationV2 result, f) AblationV3 result.

c) Representational Similarity Via CKA

Whereas UMAP reveals the overall shape of each encoder's latent space, CKA focuses on pairwise representational similarity between features. Specifically, we compare the features extracted by TraResNet and those extracted by Shallower to see if they learn redundant or complementary features. We then measure how alike they encode the same set of inputs. We find low CKA between and across all occlusion bins, indicating each level 0 model captures non-overlapping and diverse representations. Those results are: CKA_o(TraResNet, Shallower) = 0.09, CKA_1(TraResNet, Shallower) = 0.103, CKA_2(TraResNet, Shallower) = 0.2, CKA_3(TraResNet, Shallower) = 0.2273, CKA_4(TraResNet, Shallower) = 0.2248, CKA_5(TraResNet, Shallower) = 0.34, where CKA_i means the CKA value in the i-th bin.

D. Combining S_{OCL}NET and I_{FACE}NET

The *I_{FACE}NET* reconstruction model was trained to reconstruct an input image containing a face using contextual information. To guide the model in using the correct contextual information, *I_{FACE}NET* was fine-tuned using *S_{OCL}NET*'s outputs as training images that are compared in the loss function (11) with its corresponding face image. The fine-tuning is carried out for 3 epochs with the Adam optimizer, a learning rate of 0.00001, and a batch size of 8. With fine-tuning, *I_{FACE}NET* can reconstruct occluded faces. The results shown in Fig. 9. Based on this, it is proven that *I_{FACE}NET* can solve the problem of reconstructing faces occluded by an occluding object because it successfully reconstructs the occluded regions obtained by *S_{OCL}NET*.

E. Self-supervised Approach for Inpainting

We conducted additional experiments in which we trained our architecture *I_{FACE}NET* from scratch using a self-supervised strategy that augments the diversity of occlusion objects. In this setup, we removed all region-prior knowledge the binary masks gave. Now, the model must learn which regions are occluded and proceed to inpainting those regions. We presented to *I_{FACE}NET* partially or highly occluded images, allowing the

TABLE IV
METRICS OF ABLATION STUDIES. ANALYSIS OF MANIFOLD DISTRIBUTION ACROSS ALL BINS USING UMAP

Models	Metrics					
	Silhouette score		Davies-Bouldin Index		PCA Variance	
	mean	std	mean	std	mean	std
IFACENET2	**-0.049**	**±0.00311**	**8.036**	**±0.699**	**0.0184**	**±0.0018**
AblationV1	-0.0095	±0.01	6.8516	±1.55	3.9594	±1.58
AblationV2	-0.059	±0.043	3.9594	±1.58	0.021	±0.001
AblationV3	-0.0102	±0.02	7.7859	±1.309	0.0265	±0.005

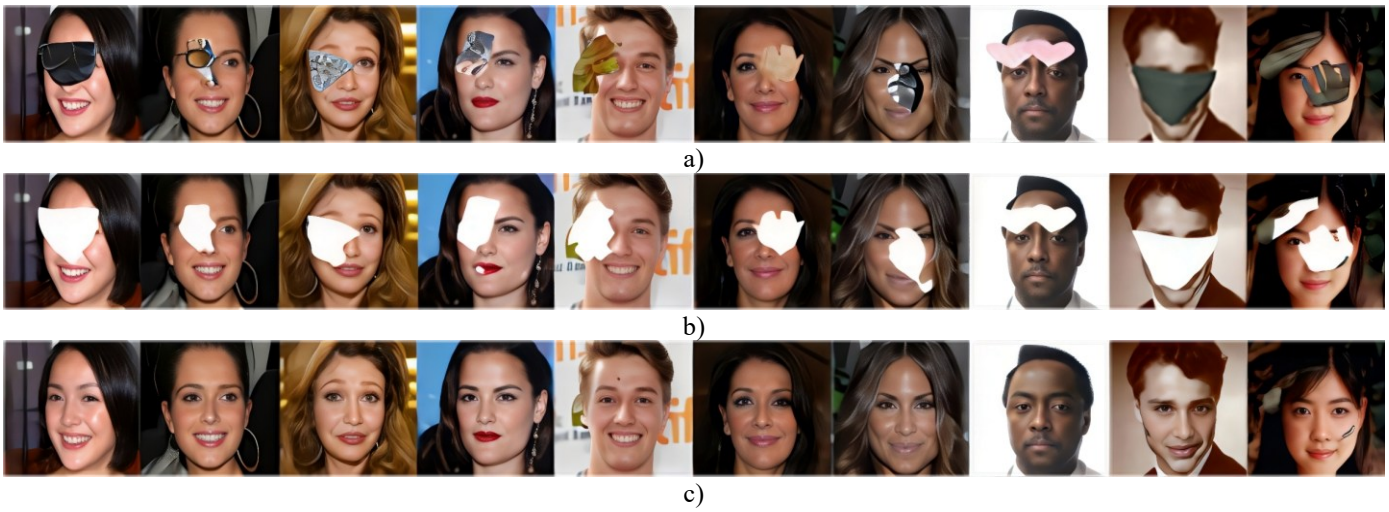


Fig. 9. Other examples show the model's ability to reconstruct occluded faces using improved contextual information. a) Image with occluded face, I_{inc}, b) output of S_{OCL}NET, I_m, c) Reconstructed face using I_{FACE}NET, I'.

model to learn how to reconstruct these images without explicit guidance on where occlusions occur. We retained all the hyperparameters from our original setup. Two types of occluders are introduced: 1) real occlusor objects from RealOcc dataset and 2) randomly generated occlusion patterns of varying shapes and sizes. This combination significantly expanded the diversity of occlusions encountered during training. Our quantitative results are MAE = 0.0412, PSNR = 24.43, SSIM = 0.94, LPIPS = 0.0024, and FID = 0.2324, and the qualitative results are shown in Fig. 10; both quantitative and qualitative results are from the validation set. These results suggest that our model can also handle occluded face inpainting with self-supervised learning.

F. General Inpainting Tasks

There is a potential risk of overfitting when employing our ensemble model *IFACENET*, which has a relatively large number of parameters. We conducted additional experiments on an entirely different dataset with considerably fewer images to address this concern and *IFACENET* capacities for solving general inpainting tasks beyond face inpainting. To this end, we trained *IFACENET* with the STL-10 dataset, which contains 8000 images categorized into 10 distinct classes. We split the dataset into 7000 images for training and 1000 for validation. We keep all hyperparameters the same as with the previous training. Despite the limited dataset sizes, the model exhibited stable performance and good metrics-wise performance: MAE = 0.0482, PSNR = 28.3, SSIM = 0.95, LPIPS = 0.0134, and FID = 4.66. Next, on Fig. 11, we share a sample of the inpainting results for the validation images.

V. CONCLUSION

This study introduces an innovative model for reconstructing occluded faces composed of *SOCLNET* and *IFACENET*. *SOCLNET* is a model designed to segment faces and objects. *SOCLNET* was designed to segment an object only when it acts as an occluder, as observed in Fig. 4. Additionally, an analysis was conducted using the XAI Grad-CAM method to provide qualitative validation of the model's performance. The reconstruction model, *IFACENET*, is based on an autoencoder, in which ensemble learning is implemented to enhance the diversity of extracted features and reduce the risk of overfitting. Using ensemble learning makes *IFACENET* competitive with state-of-the-art models. As shown in TABLE II *IFACENET* is the only model ranked among the top two models in all metrics. Ablation studies demonstrate that removing any sub-encoder drops the model's performance and weakens the latent space representation. Moreover, qualitative results of the model demonstrated accurate reconstruction for various types of faces and different mask sizes, as shown in Fig. 7. *IFACENET* can achieve the reconstruction of faces occluded by various occluders, as depicted in Fig. 9, with potential applications in real-world scenarios. Besides, *IFACENET* can reconstruct occluded faces when trained in a self-supervised strategy. We also trained our *IFACENET* model with a general dataset with fewer images and more classes, and the model still performs correctly.

Finally, on a single NVIDIA GeForce RTX 3060, our method achieves an inference speed of approximately 177 FPS on 256x256 images. This demonstrates the feasibility of our approach for real-time applications.

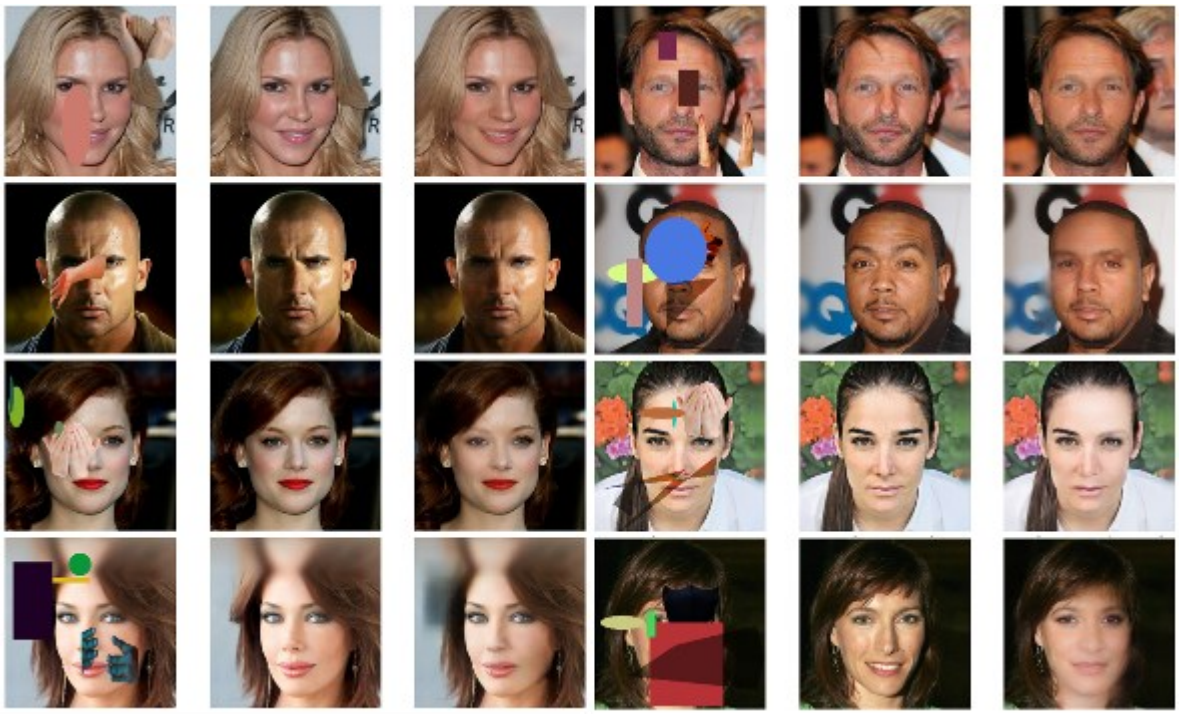


Fig. 10. Results of *I*$_{FACE}$*NET* trained with self-supervised learning approach. a) input image, b) original image, c) *I*$_{FACE}$*NET* result.

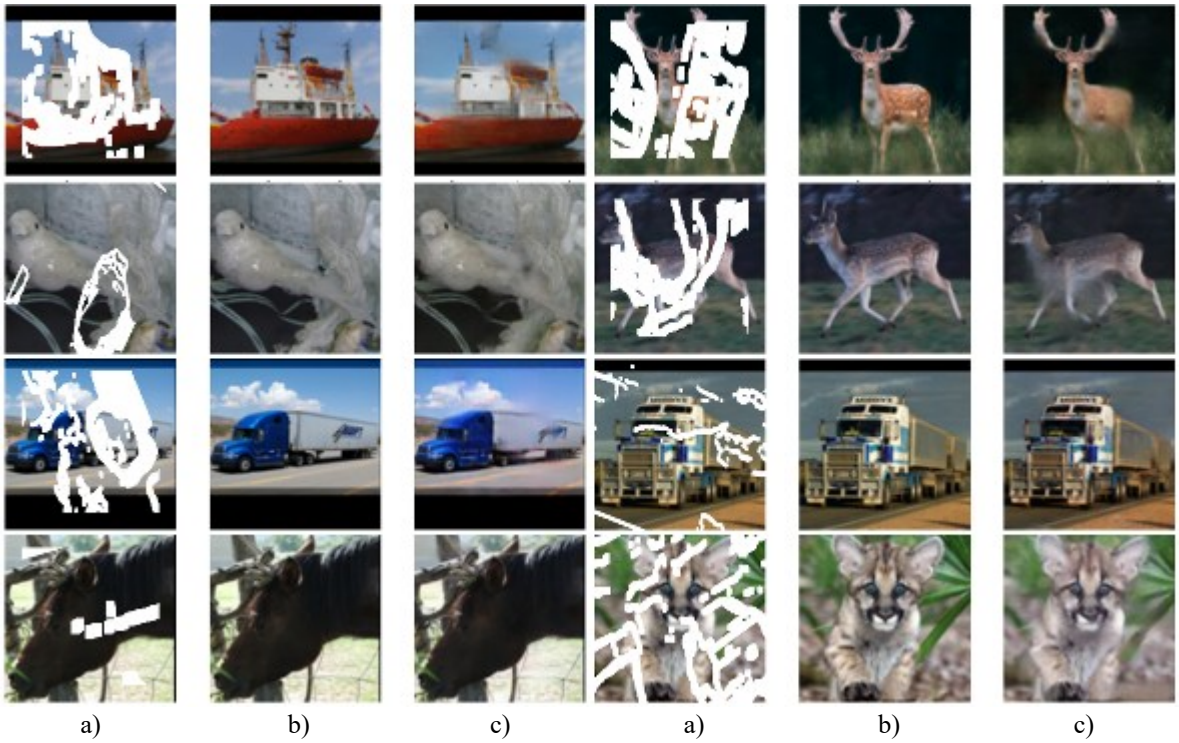


Fig. 11. Results of *I*$_{FACE}$*NET* trained with STL-10 dataset. a) input image, b) original image, c) *I*$_{FACE}$*NET* result.

REFERENCES

[1] H. Xiang *et al.*, "Deep learning for image inpainting: A survey," Pattern Recognition, vol. 134, pp. 109046, 2023. https://doi.org/10.1016/j.patcog.2022.109046

[2] F. Qin *et al.*, "Image inpainting based on deep learning: A review," Displays, vol. 69, pp. 102028, 2021. https://doi.org/10.1016/j.displa.2021.102028

[3] D. Kingma and M. Welling, "Auto-Encoding Variational Bayers," 2013. https://doi.org/10.48550/arXiv.1312.6114.

[4] H. Vega *et al.*, "Comparative study of methods to obtain the number of hidden neurons of an auto-encoder in a high-

dimensionality context," IEEE Latin America Transactions, vol. 18, no. 12, pp. 2196-2203, 2020. DOI: 10.1109/TLA.2020.9400448.

[5] I. Goodfellow *et al.*, "Generative Adversarial Networks," *Advances in neural information processing systems*, vol. 27, 2014.

[6] X. Yu *et al*., "AGG: attention-based gated convolutional GAN with prior guidance for image inpainting," Neural Computing and Applications, pp. 1-17, 2024. https://doi.org/10.1007/s00521-024-09785-w.

[7] D. Pathak *et al*., "Context Encoders: Feature Learning by Inpainting," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2536-2544, 2016. DOI: 10.1109/CVPR.2016.278.

[8] S. Iizuka *et al*., "Globally and locally consistent image completion," ACM Transactions on Graphics., vol. 36, no. 4, pp. 1–14, 2017. https://doi.org/10.1145/3072959.307365.

[9] U. Demir and G. Unal, "Patch-Based Image Inpainting with Generative Adversarial Networks," arXiv preprint arXiv:1803.07422, 2018.

[10] J. Yu *et al*., "Generative Image Inpainting with Contextual Attention," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5505–5514, 2018. DOI: 10.1109/CVPR.2018.00577.

[11] W. Quan *et al*., "Image Inpainting With Local and Global Refinement," IEEE Transactions on Image Processing, vol. 31, pp. 2405–2420, 2022. DOI: 10.1109/TIP.2022.3152624.

[12] Y. Dogan and H. Y. Keles, "Iterative facial image inpainting based on an encoder-generator architecture," Neural Computing and Applications, vol. 34, no. 12, pp. 10001–10021, 2022. https://doi.org/10.1007/s00521-022-06987-y.

[13] J. Yu *et al*., "Reference-guided face inpainting with reference attention network," Neural Computing and Applications, vol. 34, no. 12, pp. 9717–9731, 2022. https://doi.org/10.1007/s00521-022-06961-8.

[14] P. Shamsolmoali *et al*., "TransInpaint: Transformer-based Image Inpainting with Context Adaptation," in *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 849–858, 2023. DOI: 10.1109/ICCVW60793.2023.00092.

[15] S. S. Phutke *et al*., "Blind Image Inpainting via Omni-dimensional Gated Attention and Wavelet Queries," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1251–1260, 2023.

[16] X. Wang *et al*., "Spatially adaptive multi-scale contextual attention for image inpainting," Multimedia Tools and Applications, vol. 81, no. 22, pp. 31831–31846, 2022. https://doi.org/10.1007/s11042-022-12489-9.

[17] J. Jiang *et al*., "Parallel adaptive guidance network for image inpainting," Applied Intelligence, vol. 53, no. 1, pp. 1162–1179, 2023. https://doi.org/10.1007/s10489-022-03387-6.

[18] D. Cha *et al*., "SAC-GAN: Face Image Inpainting with Spatial-Aware Attribute Controllable GAN," in *16th Asian Conference on Computer Vision*, pp. 202–218, 2023. https://doi.org/10.1007/978-3-031-26293-7_13

[19] Y. Zeng *et al*., "Aggregated Contextual Transformations for High-Resolution Image Inpainting," IEEE Transactions on Visualization and Computer Graphics, vol. 29, no. 7, pp. 3266–3280, 2023. DOI: 10.1109/TVCG.2022.3156949

[20] X. Ma *et al*., "A Novel Generative Image Inpainting Model with Dense Gated Convolutional Network," International journal of computers communications & control, vol. 18, no. 2, 2023. https://doi.org/10.15837/ijccc.2023.2.5088.

[21] Y. Chen *et al*., "RNON: image inpainting via repair network and optimization network," International Journal of Machine Learning and Cybernetics, vol. 14, no. 9, pp. 2945–2961, 2023. https://doi.org/10.1007/s13042-023-01811-y.

[22] J. Wang *et al*., "Self-Prior Guided Pixel Adversarial Networks for Blind Image Inpainting," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 10, pp. 12377–12393, 2023. DOI: 10.1109/TPAMI.2023.3284431

[23] A. Mohammed and R. Kora, "A comprehensive review on ensemble deep learning: Opportunities and challenges," Journal of King Saud University-Computer and Information Sciences, vol. 35, no. 2, pp. 757–774. 2023. https://doi.org/10.1016/j.jksuci.2023.01.014

[24] J. Jam *et al*., "A comprehensive review of past and present image inpainting methods," Computer vision and image understanding, vol. 203, p. 103147, 2021. https://doi.org/10.1016/j.cviu.2020.103147

[25] O. Elharrouss *et al*., "Image Inpainting: A Review," Neural Processing Letters, vol. 51, no. 2, pp. 2007–2028, 2020.

[26] L. Trevisan de Souza *et al*., "A review on Generative Adversarial Networks for image generation," Computers & Graphics, vol. 114, pp. 13–25, 2023. https://doi.org/10.1016/j.cag.2023.05.010

[27] X. Zhang *et al*., "Image inpainting based on deep learning: A review," Information Fusion, vol. 90, pp. 74–94, 2023. https://doi.org/10.1016/j.inffus.2022.08.033.

[28] W. Quan *et al*., "Deep Learning-Based Image and Video Inpainting: A Survey," International Journal of Computer Vision, 2024. https://doi.org/10.1007/s11263-023-01977-6.

[29] Z. Yan *et al*., "Shift-Net: Image Inpainting via Deep Feature Rearrangement," in *Proceedings of the European conference on computer vision (ECCV)*, 2018. https://doi.org/10.1007/978-3-030-01264-9_1

[30] G. Liu *et al*., "Image Inpainting for Irregular Holes Using Partial Convolutions," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 85–100, 2018. https://doi.org/10.1007/978-3-030-01252-6_6

[31] C. Xie *et al*., "Image Inpainting With Learnable Bidirectional Attention Maps," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8857–8866, 2019. DOI: 10.1109/ICCV.2019.00895

[32] C. Chen *et al*., "Rethinking Atrous Convolution for Semantic Image Segmentation," 2017.

[33] Y. Zhang *et al*., "Art image inpainting via embedding multiple attention dilated convolutions," Multimedia Tools and Applications, vol. 83, no. 12, pp. 36455–36468, 2023. https://doi.org/10.1007/s11042-023-15285-1

[34] H. Liu *et al*., "Rethinking Image Inpainting via a Mutual Encoder-Decoder with Feature Equalizations," in *16th European Conference on Computer Vision*, pp. 725–741, 2020. https://doi.org/10.1007/978-3-030-58536-5_4

[35] J. Qin *et al*., "Multi-scale attention network for image inpainting," Computer Vision and Image Understanding, vol. 204, p. 103155, 2021. https://doi.org/10.1016/j.cviu.2020.103155

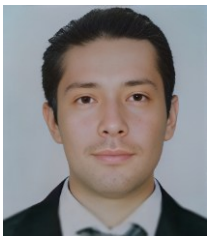
[36] Y. Chen *et al*., "DARGS: Image inpainting algorithm via deep attention residuals group and semantics," Journal of King Saud University-Computer and Information Sciences, vol. 35, no. 6, p. 101567, 2023. https://doi.org/10.1016/j.jksuci.2023.101567

[37] S. Ge *et al*., "Occluded Face Recognition in the Wild by Identity-Diversity Inpainting," IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 10, pp. 3387–3397, 2020. DOI: 10.1109/TCSVT.2020.2967754

[38] H. Li *et al*., "Recovery-Based Occluded Face Recognition by Identity-Guided Inpainting," Sensors, vol. 24, no. 2, p. 394, 2024. https://doi.org/10.3390/s24020394

[39] C. Li *et al*., "Occluded Face Recognition by Identity-Preserving Inpainting," Cognitive Internet of Things: Frameworks, Tools

and Applications, pp. 427–437, 2020. https://doi.org/10.1007/978-3-030-04946-1_41
[40] Y. Yang *et al*., "Generative face inpainting hashing for occluded face retrieval," International Journal of Machine Learning and Cybernetics, vol. 14, no. 5, pp. 1725–1738, 2023. https://doi.org/10.1007/s13042-022-01723-3
[41] A. Chen *et al*., "Occlusion-aware face inpainting via generative adversarial networks," in *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 1202–1206, 2017. DOI: 10.1109/ICIP.2017.8296472
[42] W. Jiang *et al*., "A new occluded face recognition framework with combination of both Deocclusion and feature filtering methods," Multimedia Tools and Applications, vol. 81, no. 23, pp. 33867–33896, 2022. https://doi.org/10.1007/s11042-022-12851-x
[43] X. Yuan and I. K. Park, "Face De-Occlusion Using 3D Morphable Model and Generative Adversarial Network," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10061–10070, 2019. DOI: 10.1109/ICCV.2019.01016
[44] Z. Li *et al*., "Face Inpainting via Nested Generative Adversarial Networks," IEEE Access, vol. 7, pp. 155462–155471, 2019. DOI: 10.1109/ACCESS.2019.2949614
[45] D. Kim and U. Park, "Guidance Information Assisted Reconstruction of Masked Faces," IEEE Access, vol. 11, pp. 97014–97023, 2023. DOI: 10.1109/ACCESS.2023.3311717
[46] I. Lee *et al*., "Latent-OFER: Detect, Mask, and Reconstruct with Latent Vectors for Occluded Facial Expression Recognition," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1536–1546, 2023. DOI: 10.1109/ICCV51070.2023.00148
[47] L. Li *et al*., "Mask-FPAN: Semi-supervised face parsing in the wild with de-occlusion and UV GAN," Computers & Graphics, vol. 116, pp. 185–193, 2023. https://doi.org/10.1016/j.cag.2023.08.003
[48] J. Xu *et al*., "Personalized Face Inpainting with Diffusion Models by Parallel Visual Attention," in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5420–5430, 2024. DOI: 10.1109/WACV57701.2024.00535
[49] X. Yin *et al*., "Segmentation-Reconstruction-Guided Facial Image De-occlusion," in *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–8, 2023. DOI: 10.1109/FG57933.2023.10042570
[50] K. He *et al*., "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. DOI: 10.1109/CVPR.2016.90
[51] T. Lin *et al.*, "A survey of transformers," AI Open, vol. 3, pp. 111–132, 2022. https://doi.org/10.1016/j.aiopen.2022.10.001
[52] O. K. Oyedotun, K. Al Ismaeil, and D. Aouada, "Training very deep neural networks: Rethinking the role of skip connections," Neurocomputing, vol. 441, pp. 105–117, 2021. https://doi.org/10.1016/j.neucom.2021.02.004
[53] A. E. Orhan and X. Pitkow, "Skip Connections Eliminate Singularities," arXiv preprint arXiv:1701.09175, 2017.
[54] H. Li, "Visualizing the Loss Landscape of Neural Nets," in *Advances in Neural Information Processing Systems 31 (NIPS 2018)*, 2017.
[55] G. Lu *et al*., "Optimizing Depthwise Separable Convolution Operations on GPUs," IEEE Transactions on Parallel and Distributed Systems, vol. 33, no. 1, pp. 70–87, 2022. DOI: 10.1109/TPDS.2021.3084813
[56] J. Johnson *et al*., "Perceptual Losses for Real-Time Style Transfer and Super-Resolution," in *Computer Vision - ECCV 2016*, pp. 694–711, 2016. https://doi.org/10.48550/arXiv.1603.08155
[57] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv, 2014. https://doi.org/10.48550/arXiv.1409.1556.
[58] T. Akiba *et al.*, "Optuna: A next-generation hyperparameters optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 2623-2631, 2019. https://doi.org/10.1145/3292500.333070.
[59] Y. Ozaki *et al.*, "Multiobjective tree-structured parzen estimator for computationally expensive optimization problems," in *Proceedings of the 2020 genetic and evolutionary computation conference,* pp. 533-541, 2020. https://doi.org/10.1145/3377930.3389817.
[60] R. Voo *et al*., "Delving into High-Quality Synthetic Face Occlusion Segmentation Datasets," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4710–4719, 2022. DOI: 10.1109/CVPRW56347.2022.00517
[61] R. Geirhos *et al*., "Shortcut learning in deep neural networks," Nature Machine Intelligence, vol. 2, no. 11, pp. 665–673, 2020. https://doi.org/10.1038/s42256-020-00257-z
[62] X. Bai *et al*., "Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments," Pattern Recognition, vol. 120, p. 108102, 2021. https://doi.org/10.1016/j.patcog.2021.108102
[63] R. Selvaraju *et al*., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017. DOI: 10.1109/ICCV.2017.74
[64] T. Huang, "IntroVAE: Introspective Variational Autoencoders for Photographic Image Synthesis," in *Advances in neural information processing systems*, vol. 31, 2018. https://doi.org/10.48550/arXiv.1807.06358
[65] J. Yu *et al*., "Free-Form Image Inpainting With Gated Convolution," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4470–4479, 2019. DOI: 10.1109/ICCV.2019.00457
[66] Z. Yi *et al*., "Contextual Residual Aggregation for Ultra High-Resolution Image Inpainting," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7505–7514, 2020. DOI: 10.1109/CVPR42600.2020.00753
[67] L. McInnes *et al.*, "UMAP: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018. https://doi.org/10.48550/arXiv.1802.03426.
[68] S. Kornblith *et al*., "Similarity of neural network representations revisited," in *International Conference on Machine Learning*, pp. 3519-3529.



Miguel A. Gutierrez-Velazquez received the B.Sc., and M.Sc. in Electronic Engineering in 2019 and 2021, respectively from the Instituto Tecnologico Nacional de Mexico campus I.T. Chihuahua.

Mario I. Chacon-Murguia received the B.Sc. and M.Sc. degrees in EE from the Chihuahua Institute of Technology, Mexico, in 1982 and 1985, respectively, and the Ph.D. degree in EE from New Mexico State University, USA, in 1998. He has developed research projects for several companies. He is currently a Research Professor with the Chihuahua Institute of Technology, and the Director of the Visual Perception Laboratory. He has published more than 175 works and published three books. His current research includes computer vision, and image and signal processing using computational intelligence. He is a member of the IEEE Computational Intelligence Society, the IEEE Digital Signal Processing, and the National Research System in Mexico.

Juan A. Ramirez-Quintana received the B.Sc., M.Sc., and Ph.D degrees in Electronic Engineering in 2004, 2007, and 2014, respectively. From 2008 to 2011, he was a researcher and a teaching assistant at different universities. He is currently a Researcher and Professor at the Tecnologico Nacional de Mexico campus I.T. Chihuahua and the Director of the Digital Signal Processing and Artificial Intelligence Lab. His current research includes computer vision, signal processing, computational intelligence, and machine learning. He is the author of patents and more than 50 papers published in journals, proceedings, and book chapters. He is a member of the National Research System of Mexico (SNII) and the Academia Mexicana de Computación (AMEXCOMP).