Simulating Real-World Challenges: Blind Face Restoration and Upscaling

Elena Zoppellari
elena.zoppellari@studenti.unipd.it

Rodrigo Golan
rodrigo.golan@studenti.unipd.it

Abstract

Blind Face Restoration tackles the intricate challenge of reconstructing high-quality face images from low-quality counterparts, sans knowledge of degradation specifics. This study simulates blind face images by applying diverse degradations to the Labeled Faces in the Wild (LFW) dataset. This work adopts Kim et al.'s progressive 3-step architecture[8], initially designed for Face Super Resolution, and further develops it through the design of an attention loss utilizing facial landmarks for two diverging tasks: 16x16 blind face images restoration and upscaling, simulating challenging real-life scenarios, and standard face reconstruction from blind face inputs. Dynamic weighting is exploited to prioritize stability or attention to facial details and perceptual fidelity during training. Albeit not always visually satisfactory individually, the outcomes of our models might suggest a potentially viable approach for images with multiple levels of proximity which could benefit from a synergic collaboration of two such networks for enhanced blind face restoration.

1. Introduction

Blind Face Restoration, a specialized task within the broader domain of face restoration, involves reconstructing high-quality face images from low-quality counterparts without knowledge of degradation types or parameters. This challenging problem finds applications in face recognition, law enforcement, and privacy protection. In this study, blind face images are simulated by applying a random combination of noise, blur, low resolution, and JPEG compression artifacts to the Labeled Faces in the Wild [7] dataset.

Within blind face reconstruction, methods diverge into two categories [16]: prior-based deep restoration and non-prior-based deep learning. While some incorporate facial prior knowledge, others aim to map low-quality to high-quality facial images without such priors. The former category further branches into three sets—geometric, reference, and generative priors—each delineating distinct methodologies. Cutting-edge networks in this domain integrate facial priors, pre-trained GAN models, and Vision Transformer architectures.

Figure 1. Objective 1: Blind Face Restoration and upsampling

Figure 2. Objective 2: Blind Face Restoration

This work adopts Kim et al.'s progressive 3-step architecture [8] as a starting point, initially designed for Face Super Resolution, developing it for Blind Face Restoration. In particular, we make use of facial landmarks[3] as geometrical priors to enhance facial restoration and employ dynamic weighting for hierarchical feature learning. Two distinct objectives are pursued: upscaling 16x16 blind face images to 128x128 (O1) and standard face reconstruction from blind inputs at 128x128 (O2).

To emphasize the practical relevance of O1, we draw parallels with real-world scenarios involving distant faces or other challenging conditions, such as group photos, both leading to low-resolution images. Thus, this objective simulates situations where facial details are harder to discern,

aligning the research with possibly common real world applications. For O2 which lies in a more standard Blind Face Restoration framework with no upscaling, we design and incorporate an encoder. In order to focus on facial feature reconstruction we exploit perceptual and attention losses, and WGAN loss with gradient penalty.

Evaluation metrics include PSNR, SSIM, MS-SSIM, LPIPS, NIQE, and FID. The results demonstrate partial success of the proposed networks in generating high-fidelity face images for both objectives.

In summary, our contributions include:

1. Objective Diversity: We introduce two distinct objectives—(O1) upscaling 16x16 blind face images to 128x128, simulating scenarios with distant or challenging conditions , and (O2) standard face reconstruction at 128x128.

2. Architectural Enhancements: Adapting Kim et al.'s progressive 3-step architecture to a different task, we integrate facial landmarks and introduce an encoder.

3. Effective Model Training: Leveraging perceptual and attention losses, we cautiosly employ dynamic weighting balancing between stability and increased emphasis on the attention to facial details and perceptual fidelity.

2. Related Work

In the domain of face restoration, various approaches have been explored, ranging from direct Convolutional Neural Network (CNN) models such as Zhou et al. [1], to more prevalent generative models due to their capability to generate visually appealing images. Recently, with the successes of attention mechanisms and Transformer architectures [15], experiments have been conducted using Vision Transformer (ViT) [4] and its variants, as seen in the notable case of FaceFormer [9].

Surveyed by Wang et al. (2022) [16] and Li et al. (2023) [10], the prevalent technique for enhancing face restoration involves incorporating a prior into the architecture. This prior can take the form of additional information about the face being restored, as demonstrated by Liu et al. [12]. An advancement of this approach are facial dictionaries, consisting of facial elements categorized from a high-quality face dataset that the network progressivly chooses during training. As example, this technique was implemented by Li et al. in their work [11].

Another approach involves geometrical facial priors, represented by landmarks, heatmaps (as implemented by Kim et al. [8]), semantic maps, or 3D face priors (as implemented by Hu et al. [6]). A combination of different geometric priors has also been proposed by Yu et al. [19]. The power of pre-trained generative priors has also been extensively explored: they can automatically extract information beyond facial features, including texture and hair details, making approaches based on pre-trained generative priors simpler and more efficient. State-of-the-art models, such as GPEN proposed by Yang et al. [18] and GFPGAN proposed by Wang et al. [17], have demonstrated the efficacy of these approaches.

The challenge we're tackling is quite peculiar: blind face restoration is a complex problem due to its ill-posed nature, which arises from the unknown degradation process. This sets it apart from non-blind tasks where degradation parameters remain constant. To gain some insights into this intricate problem, considering our limited computational resources, we opted for a simpler architecture compared to the state-of-the-art models.

It is worth mentioning that to extract facial landmarks, crucial in the attention loss estimation, we employed Bulat et al.'s Face Alignment Network [3]. This network introduces a groundbreaking baseline, combining a state-of-the-art landmark localization architecture with a powerful residual block. Trained on a large synthetically expanded 2D facial landmark dataset, the model is evaluated across various other 2D facial landmark datasets, marking a notable advancement in the field.

3. Dataset

The dataset chosen for this study is *Labeled Faces in the Wild* by Huang et al. (2008) [7], comprising 13,233 unaligned images collected from the web. The selection of this dataset adds an additional layer of challenge to the task, as the images lack systematic alignment for landmarks identification and face reconstruction. Originally intended for face recognition tasks, we opted for this dataset due to its faithful representation of authentic facial features and its relatively compact size, which aligns with our computational resources. Our expectation is that training the model on a more realistic set of images will enhance its utility in practical scenarios.

LFW images are transformed into *blind* images, that means images that had undergone an unknown degradation process. This process could involve a combination of various fundamental processes considered in image reconstruction, including blurring (representing, for example, the effects of poor focus in a photograph), low quality (resulting from resizing larger images), noise (causing disturbances in pixel intensities), and artifacts (common in lossy compression methods such as JPEG compression). Blind images are built considering a random combination of these degradation techniques, aiming to produce realistically deteriorated images and preventing the model from learning to solve only one sectorial task. In fact, the Blind Face Restoration (BFR) task aims to build an algorithm capable of recon-

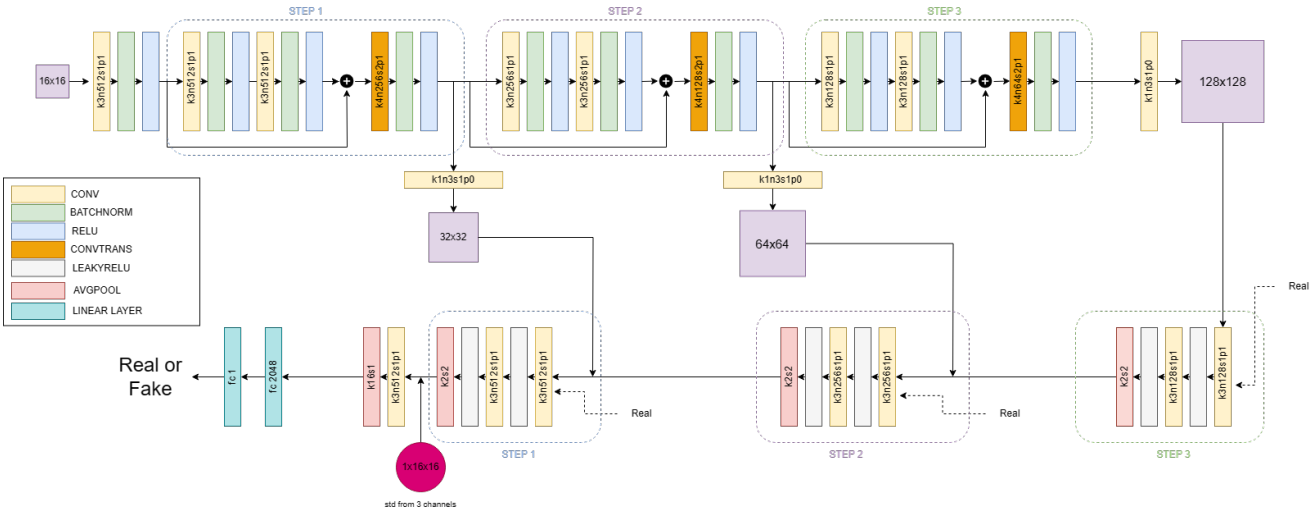

Figure 3. Progressive GAN architecture

structing human faces regardless of the specific degradation process they underwent, making it more suitable for realistic face restoration.

The generation of blind images followed the methodology outlined by Li et al. in their 2019 work [11]. This multistage process began with the convolution of the original image using a blurring kernel, followed by downsampling. Subsequently, noise was added to the image, followed by conversion to the JPEG format. Finally, the image was upsampled to achieve the input dimensions, different for the two mentioned objectives, and pixel values were normalized to the [0,1] range. This entire process can be summarized by the equation:

$$I_{blind} = JPEG_q[(I_{orig} \star k_\sigma) \downarrow_{s\times s} + n_\delta] \uparrow_{u\times u} \quad (1)$$

As mentioned in the introduction, we developed two architectures for our tasks: one for the blind reconstruction of 16×16 images and another for 128×128 images. Consequently, the dataset underwent two slightly distinct preprocessing procedures employing a custom data loader tailored to generate degraded images from the originals. Initially, both preprocessing steps involved resizing LFW images to the desired target dimensions of $128 \times 128 \times 3$ (creating in this way the true labels). At this point, equation 1 is applied to the images, choosing the parameters in the following ranges:

- $u \times u = 16 \times 16$ imges: $k \in [1, 5]$, $\sigma \in [0, 5]$, $s \in [43, 128]$, $\delta \in [0, 10]$, and $q \in [60, 100]$
- $u \times u = 128 \times 128$ images: $k \in [1, 11]$, $\sigma \in [0, 10]$, $s \in [16, 128]$, $\delta \in [0, 30]$, and $q \in [50, 100]$

where k is the blurring kernel, σ is the blurring intensity, s represents downsampling new size, δ is the noise level and q indicates JPEG quality. These ranges were selected differently in order to allow 16×16 images to exhibit an amount of blindness suitable for their small size. The level of degradation is calculated using PSNR, SSIM, MS-SSIM, LPIPS, NIQE and FID metrics for both the two sets of blind images and reported in Table 1 and Table 2.

blind img	PSNR↑	SSIM↑	MS-SSIM↑
$16x16$	20.1 ± 0.2	0.59 ± 0.01	0.818 ± 0.002
$128x128$	21.1 ± 0.7	0.57 ± 0.03	0.83 ± 0.02

Table 1. Batch metrics for input blind images (1)

blind img	LPIPS↓	NIQE↓	FID↓
$16x16$	0.468 ± 0.003	16.78 ± 0.05	240 ± 4
$128x128$	0.52 ± 0.03	9.8 ± 0.5	151 ± 14

Table 2. Batch metrics for input blind images (2)

At this point, for each image, the true labels are downsampled by $2\times$ and $4\times$ and collected. As aforementioned, one of our objectives is to use facial landmarks as geometrical priors to enhance facial restoration. To achieve this, we extract the facial landmarks from the true label image and its corresponding $2\times$ downsample, using the pre-trained Face Alignment Network (FAN) by Bulat and Tzimiropoulos [3]. Since the number of landmarks per image vary, in this preprocessing phase, they are inserted into tensors with the same dimensions to facilitate retrieval during training.

Finally, the preprocessed images are divided into train and test sets, comprising individually 10,586 and 2,650 images. For the 16×16 blind images, the batch size is set to 32, while for the 32×32 blind images, it is set to 16.

4. Method

4.1. Architecture

As previously introduced, we have employed two architectures to address the proposed tasks. The first one is a Generative network strongly inspired by Kim et al. (2019) [8], where direct facial landmarks have been included instead of facial heatmaps. The second architecture is an extension of the first one, with the goal of adapting the task to larger input images, employing an encoder.

4.1.1 Generator and Discriminator

The progressive GAN architecture is illustrated in Figure 3, consisting of a Generator (upper figure) and a Discriminator (lower figure). Initially proposed for the Super Resolution (SR) task, we have adapted this model to address the challenge of restoring small blind images.

The model employs progressive training, where each batch undergoes successive steps. Initially, the images pass only through step 1 of the generator, creating a 32×32 image that is then processed by step 1 of the discriminator. Then, the same batch progresses through step 1 and step 2 of the generator, generating a 64×64 image that undergoes step 2 and step 1 of the discriminator. Finally, all images in the batch traverse all three steps of the generator, resulting in 128×128 images that are processed by all steps of the discriminator. This progressive training strategy enables the model to better learn reconstruction at each resolution, aiming for higher-quality restored images.

In Figure 3, each step of the generator includes a residual block, a transposed convolutional layer, a batch normalization layer and a ReLU activation. The residual block consists of two convolutional layers followed by a batch normalization layer and a ReLU activation function. The transposed convolutional layer expands the image size by a factor of 2 while reducing the channel dimension. For each step, the generated image passes through a convolutional layer to reconstruct the three RGB channels before being forwarded to the respective discriminator step.

In each step of the discriminator, two convolutional layers increase the feature dimension, followed by a leaky ReLU activation and an average pooling layer that reduces the image size by a factor of 2. After the first step, the discriminator concatenates the standard deviation along the channels of the input $3 \times 16 \times 16$ image to the resulting $512 \times 16 \times 16$ tensor, creating a $513 \times 16 \times 16$ tensor. This tensor then undergoes a convolutional layer, an average pooling layer, and two linear layers, which first expand the features and then produce the final probability value for each element in the batch.

4.1.2 Encoder

To address blind 128×128 images, an encoder has been integrated into the Progressive GAN architecture. The idea is to gradually downsample the input image while increasing the feature dimension, with the goal of generating a $512 \times 16 \times 16$ tensor. This tensor is then fed into step 1 of the GAN, containing more information about the original image compared to straightforwardly downsampling the input size. Additionally, this approach aims to generalize the architecture to handle tasks beyond Super Resolution (SR).

The encoder, illustrated in Figure 4, consists of three downsampling blocks. Each block comprises two convolutional layers followed by a leaky ReLU activation function and a batch normalization layer, which is subsequently followed by another ReLU activation function. The first convolutional layer in each block is responsible for downsampling the image by a factor of 2 while simultaneously increasing the feature dimension.

4.2. Losses

A multifaceted loss strategy is employed to collectively guide the training process through a delicate balance between stability and the generation of visually compelling results.

Adversarial loss : To enhance the stability of the training process, we employ the Wasserstein Generative Adversarial Network (WGAN) loss with a gradient penalty. WGAN, proposed in [2], utilizes Wasserstein's distance to minimize the discrepancy between the distribution of true target images $I_{orig} \sim P_r$ and the distribution of generated images $I_{blind} \sim P_g$:

$$W(P_r, P_g) = \max_{||f|| \leq 1} \{ \mathop{\mathbb{E}}_{x \sim P_r} [f(x)] - \mathop{\mathbb{E}}_{x \sim P_g} [f(x)] \} \quad (2)$$

where f is 1-Lipschitz continuous. The gradient penalty (GP) [5] introduces a constraint on the gradient norm of the discriminator to enforce Lipschitz continuity. This ensures more stable training of the network compared to the standard WGAN and requires minimal hyperparameter tuning. The resulting loss is given by:

$$L_{adv} = \mathop{\mathbb{E}}_{I_{orig} \sim P_r} [D(I_{orig})] - \mathop{\mathbb{E}}_{I_{blind} \sim P_g} [D(I_{blind})] + \lambda \mathop{\mathbb{E}}_{\hat{I} \sim P_{I_{blind}}} [||\nabla_{\hat{I}} D(\hat{I})_2 - 1||^2] \quad (3)$$

where $P_{\hat{I}}$ is the distribution obtained by uniformly sampling along a straight line between the real and generated distributions P_r and P_g.

MAE loss: The adopted pixel wise loss is the Mean-Absolute-Error, aiming to minimize the $L1$ distance between the generated and the target images with a more ro-

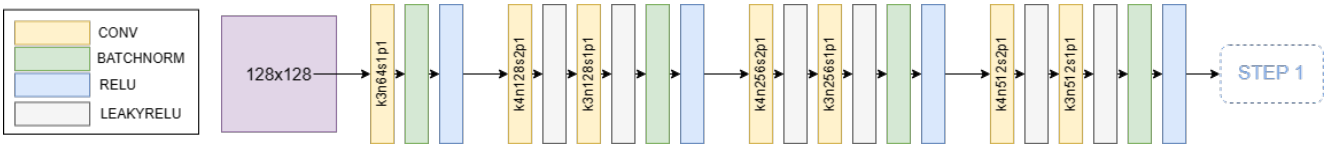

Figure 4. encoder attached to GAN architecture

bust error metric.

$$L_{pixel} = \frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H}\left|(I_{orig})_{x,y} - (\,G(I_{blind})\,)_{x,y}\right|, \quad (4)$$

Perceptual loss: We utilize the pre-trained VGG19 [13] to extract high-level semantic information from its intermediate layers. By estimating the $L1$ distance between the original and the target features, we try to ensure accurate capture of complex facial details promoting the creation of photo-realistic face images. The perceptual loss is given by:

$$L_{percep} = \sum_{i}\frac{1}{W_i H_i}\times \\ \times\sum_{x=1}^{W_i}\sum_{y=1}^{H_i}\left|\phi_i(I_{orig})_{x,y} - \phi_i(\,G(I_{blind})\,)_{x,y}\right|, \quad (5)$$

where each layer i is in the predetermined list.

4.2.1 FAN-based attention loss

The Face Alignment Network (FAN) is a neural network designed for facial landmark detection. We employ the pre-trained model by Bulat et al [3] which can accurately locate key facial points, such as eyes and nose, using a deep convolutional neural network. Our task of Face Restoration greatly benefits from FAN's ability to handle variations in pose, expression, and illumination, making it robust for landmark detection even in a real-world dataset.

The facial landmarks are extracted from the original images and also their $2\times$ downsample in the preprocessing phase. We design an attention loss that takes action in the intermediate and last step of training, weighing the channel-wise mean of the pixel $L1$ distance between the target image and the generator output in the adjacent areas to the landmarks. The weights are chosen to be those of the gaussian kernels 3×3 and 5×5 for the $2\times$ downscaled and the original size images respectively, mimicking a sort of "localized convolution" taken only in the coordinates specified by the landmarks. Our facial attention loss per image is defined as:

$$L_{atten} = \sum_{\substack{(x_l,y_l),\\ l\in\Lambda}}\left(\left|I_{orig} - G(I_{blind})\right| * h_{5x5}\right)[x_l, y_l], \quad (6)$$

where Λ is the set of landmarks of image I_{orig}, each landmark l has coordinates (x_l, y_l) and h_{5x5} is the gaussian kernel. An analogous formula is used for the downscaled images. The losses per image are then averaged to yield a single value per batch for step 2 and 3.

4.2.2 Dynamic weighting and overall training loss

The first task (O1) is highly complex and thus requires great care in designing training. Stability issues are considered prioritizing the adversarial loss at the start of training, while later effort is put to obtain visually plausible results. At the same time the network developed for the less ill posed problem in (O2) benefits from an adaptive weighting approach, which allows the model to progressively get a deeper understanding of the features of interest by increasing the corresponding weights during training. The overall training loss at step k is given by:

$$L_{train} = L_{adv} + \alpha_k L_{pixel} + \beta_k L_{percep} + \gamma_k L_{atten}, \quad (7)$$

with $\gamma_1 = 0$ and each weight $\chi_k \in \{\alpha_k, \beta_k, \gamma_k\}$, is a function of the epoch, $\chi_k = \chi_k(epoch)$. The weights are increased at most linearly with each epoch after monitoring stability. The progressive nature of the model comes in play as a key aspect ensuring that each weight must not necessarily be the same across different steps; indeed we have used $\gamma_2 \neq \gamma_3$

5. Experiments

We have trained our models independently on the constructed dataset from the ground up using NVIDIA T4 Tensor Core GPU, reaching 102 epochs for both.

5.1. Metrics

To comprehensively evaluate our results, we employed six different metrics:

1. **Peak Signal-to-Noise Ratio (PSNR):** a full-reference metric measuring pixel-wise differences between the recovered and ground-truth images, emphasizing signal fidelity. According to both [16] and [10], higher PSNR indicates lower pixel-wise differences, but it may not align well with human perception. It ranges from 0 to ∞, with the goal of maximizing the result.

2. **Structural Similarity (SSIM):** another full-reference metrics that evaluate differences in brightness, contrast, and structure between images, providing a more

holistic view compared to PSNR. It ranges from 0 to 1, with the goal of maximizing the result.

3. **Multi-Scale Structural Similarity (MS-SSIM):** it is built upon SSIM by dividing the image into windows, calculating SSIM for each, and averaging to get a comprehensive assessment. Compared to SSIM, it enhances structural similarity measurement over multiple scales for a more nuanced evaluation. It ranges from 0 to 1, with the goal of maximizing the result.

4. **Learned Perceptual Image Patch Similarity (LPIPS):** it focuses on visual perceptual similarity by leveraging a learned model rather than pixel-wise differences. For our evaluation, we considered the pre-trained VGG model [13] used for perceptual loss. It returns a perceptual assessment of visual quality, capturing aspects more aligned with human perception. It ranges from 0 to 1, with the goal of minimizing the result.

5. **Natural image quality evaluator (NIQE):** a non-reference metric useful for evaluating real images without ground-truth targets. Widely used in literature to measure the naturalness of real face images, it ranges from 0 to ∞, with the goal of minimizing the result.

6. **Fréchet inception distance (FID):** it measures the difference in distribution between generated and real images using features from a pre-trained neural network (for our model, we considered the pre-trained Inception-v3 network [14]). This metric allows assessing the visual quality of generated images in terms of distributional similarity. It ranges from 0 to ∞, with the goal of minimizing the result.

The first three metrics are pixel-wise, useful for quantitatively evaluating the resolution of predictions compared to true targets but may fall short in assessing naturalness. The bottom three are examples of full-reference, semi-reference, and non-reference metrics able to evaluate the visual appeal of an image to the human eye.

5.2. Results

The most notable outcomes of both models are depicted in Figure 1 and Figure 2, while the conclusive assessments of the models on the test set are presented in Table 3 and 4. A comparison of these tables with Table 1 and 2 reveals that the two models have not exhibited a substantial increase in pixel-wise metrics. However, there is a noteworthy enhancement in visual quality and overall human eye appeal.

5.3. Critical issues

Various issues have been observed with our proposed models where performance is low. Setting aside the cases of excessively deteriorated input images, we can look for systematic causes of underperformance. Such typical motifs are valid for both our models, while clearly impacting them differently, and involve one or a combination of: (a) multiple faces, (b) difficult poses and expressions, (c) objects interference, (d) ethnic group underrepresentation and (e) gender assessement. Outside these scenarios the results are seldom examples of coarse failed convergence or highly improper face reconstruction even for the first task.

The second network adopted for (O2) reaches high visual performance overall, but lacks in reconstructing finer details such as wrinkles (f).

pred img	PSNR↑	SSIM↑	MS-SSIM↑
$16x16$	20.6 ± 0.3	0.65 ± 0.01	0.862 ± 0.004
$128x128$	22.0 ± 0.4	0.64 ± 0.01	0.862 ± 0.009

Table 3. Batch metrics for generated images (1)

pred img	LPIPS↓	NIQE↓	FID↓
16×16	0.295 ± 0.006	7.9 ± 0.3	40 ± 2
128×128	0.31 ± 0.01	6.8 ± 0.2	35 ± 2

Table 4. Batch metrics for generated images (2)

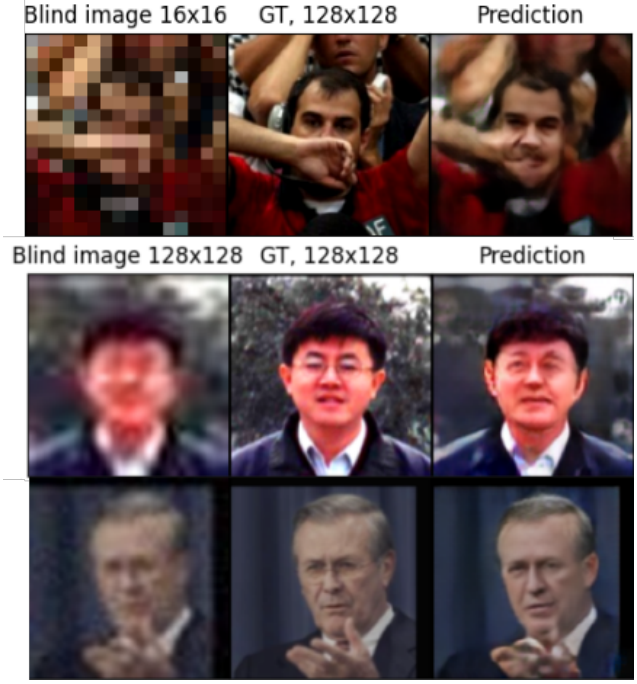

Figure 5. Examples of critical issues (from the top): (c): hand interference, (d): ethnic information lost, (f): missing fine details

6. Conclusion

This work has explored the Blind Face Restoration problem from two different angles. Our models have demonstrated sufficient ability at capturing the key features of the face, considering the demanding tasks and limited resources. Further research might focus on the combination of these networks in images with various levels of proximity and perspective in which they might be used synergistically, employing other datasets and extending the learnabale features.

With increased resources, the model could benefit from extended training time and the incorporation of a more sophisticated geometrical prior, such as semantic maps. This enhancement could effectively address the challenge of restoring highly degraded images. Additionally, the inclusion of residual blocks in the encoder architecture might further improve the model's performance.

References

[1] Learning face hallucination in the wild. 29, Mar. 2015.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017.

[3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[5] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. *CoRR*, abs/1704.00028, 2017.

[6] Xiaobin Hu, Wenqi Ren, Jiaolong Yang, Xiaochun Cao, David Wipf, Bjoern Menze, Xin Tong, and Hongbin Zha. Face restoration via plug-and-play 3d facial priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8910–8926, 2022.

[7] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[8] Deokyun Kim, Minseon Kim, Gihyun Kwon, and Dae-Shik Kim. Progressive face super-resolution via attention to facial landmark, 2019.

[9] Aijin Li, Gen Li, Lei Sun, and Xintao Wang. Faceformer: Scale-aware blind face restoration with transformers, 2022.

[10] Wenjie Li, Mei Wang, Kai Zhang, Juncheng Li, Xiaoming Li, Yuhang Zhang, Guangwei Gao, Weihong Deng, and Chia-Wen Lin. Survey on deep face restoration: From non-blind to blind and beyond, 2023.

[11] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *ECCV*, 2020.

[12] Jixin Liu, Rui Chen, Shipeng An, and Heng Zhang. Cg-gan: Class-attribute guided generative adversarial network for old photo restoration. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 5391–5399, New York, NY, USA, 2021. Association for Computing Machinery.

[13] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[14] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[16] Tao Wang, Kaihao Zhang, Xuanxi Chen, Wenhan Luo, Jiankang Deng, Tong Lu, Xiaochun Cao, Wei Liu, Hongdong Li, and Stefanos Zafeiriou. A survey of deep face restoration: Denoise, super-resolution, deblur, artifact removal, 2022.

[17] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior, 2021.

[18] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. GAN prior embedded network for blind face restoration in the wild. *CoRR*, abs/2105.06070, 2021.

[19] Yanjiang Yu, Puyang Zhang, Kaihao Zhang, Wenhan Luo, Changsheng Li, Ye Yuan, and Guoren Wang. Multi-prior learning via neural architecture search for blind face restoration, 2023.