

Generating High-Resolution Facial Images through Face Hallucination Techniques Using Deep Learning Models

ASEEL AL'NAEIMEE^{1*}, AMIR LAKIZADEH²

Abstract: In recent years, the science of computer vision has seen a significant increase in researchers interested in studying facial hallucinations. This endeavour continues to be one of the most difficult to do because of the intricate and delicate structure of the human face. In this paper, an End-to-end Generative Adversarial Network for Face Hallucination based on extending ESRGAN method is proposed. The proposed model uses Residual in Residual Dense Block (RRDBNet) in its generator to improve performance compared to ESRGAN benchmark model. The Generator creates visually appealing hallucinated face pictures. At the same time, the enhanced discriminator assesses the degree of realism in the input images. During the training phase, the mbileNet model can extract features. The suggested model will be able to conceive face representations that are perceptually credible because it will be able to have an advantage over ongoing competitive learning. Extensive studies conducted on huge face datasets demonstrate that the introduced method performs noticeably better than existing methods considered to be state-of-the-art.

INTRODUCTION

Face Hallucination (FH), sometimes known as Face Super-Resolution (FSR), is a specialized image Super-Resolution (SR) challenge. This issue pertains to generating High-Resolution (HR) facial photographs from their Low-Resolution (LR) equivalents. Face analysis is a crucial activity, significantly beneficial for computer vision applications like face surveillance [32] and recognition [21]. This work is of significant relevance. Face hallucination, however, presents an ill-posed inverse problem that is extremely challenging to resolve, as the low-resolution image may correspond to several high-resolution candidate pictures and has forfeited numerous critical face traits and components. The studies [3, 23, 24] numerous distinct face hallucination methodologies have been proposed to generate high-quality facial images. These approaches may be categorized into two types: classical and deep learning-based. A multitude of traditional methods have been proposed to tackle the issues of face hallucination. The studies [2, 15, 20] introduced the image pyramid model in their presentation, which can reconstruct high-frequency features of low-resolution facial images. This approach was developed to identify the best correlation between low-resolution and high-resolution patches. Wang and Tang [15] employed an eigen-transformation to establish a linear mapping between the low-resolution and high-resolution face subspaces. Yang, et al. [20] integrated facial priors to extract facial information from high-resolution picture components. They accomplished this by postulating a correlation among specific face elements.

¹ Computer Engineering Department, University of Qom, Qom, Iran. Email: a.alnaeimee@stu.qom.ac.ir

² Computer Engineering Department, University of Qom, Qom, Iran. Email: lakizadeh@qom.ac.ir

*Correspondence: a.alnaeimee@stu.qom.ac.ir

Numerous deep learning algorithms have recently emerged, achieving state-of-the-art performance [4, 10, 30, 31]. Dong, et al. [4] pioneered a deep learning-based super-resolution method known as SRCNN. This approach directly learns an end-to-end mapping between high-resolution images and low-resolution images. Zou et al. [10] reported a distinctive bi-channel convolutional network for generating face images derived from nature. Zhu, et al. [31] introduced the Cascaded Bi-Networks (CBN), including two sub-networks optimized alternately. The sub-networks comprised facial hallucination and dense correspondence field estimation.

To optimize the hallucination process, the approaches above employ reconstruction loss, such as L1 or L2, frequently leading to the generation of excessively smoothed hallucinated pictures. This occurs despite these models achieving a superior Peak Signal-to-Noise Ratio (PSNR) score [18]. To address this issue, many models utilizing Generative Adversarial Networks (GAN) have been proposed [3, 12, 23, 25]. Research has shown that GAN-based models with robust constraint losses may generate visually realistic high-resolution images [28]. The work [12] extended GAN into the domain of super-resolution (SR) and proposed an effective method termed SRGAN. This approach employed both an antagonistic loss and a perceptual loss. Following the SRGAN, Wang, et al. [17] introduced the ESRGAN by presenting an innovative generator architecture and utilizing enhanced perceptual loss attributes. MTDN, founded on GAN, was introduced by Yu, et al. [27]. Nonetheless, when the input resolution is exceedingly low, it fails to restore high-quality face images, leading to indistinct patterns with significant aberrations.

The GAN-based face hallucination models discussed before are susceptible to model collapse [3, 12, 17, 26], which leads to ghost artefacts in the hallucination results. This is especially true when the input picture resolution is very low. By incorporating the ESRGAN network and the MobileNet model, we present a unique hybrid GAN-based technique called end-to-end generative competitive face hallucination. This method is intended to solve the issue that has been brought to our attention. The primary contributions of this work may be broken down into three categories:

1. The approach that we have presented is capable of transforming a tiny unaligned low-resolution (LR) face picture, which is 64x64 pixels, into a hallucination (HF) face image, which is 256x256 pixels while increasing the super-zoom factor by 8x.
2. Our proposal is called Residual in Residual Dense Block (RRDBNet). It is designed to improve the flow of features via the Generator, solve the problem of gradient disappearance, and provide the hallucination process with hierarchical data.
3. The use of inverse procedures in the Generator, which increases the image's dimensions, was accomplished with the help of ConvTranspose2D.
4. Self-Attention Layers: Added after every 5th RRDB block in the Generator and after the second convolutional block in the discriminator.
5. Intermediate Feature Fusion: Introduced a feature fusion layer to combine different hierarchical features in the Generator.
6. PixelShuffle and MobileNetV2: Replacing traditional upsampling layers with PixelShuffle and MobileNetV2 Inverted Residual Blocks.
7. Discriminator Refinement: Includes ResNet-style residual blocks with added self-attention for enhanced feature learning

RELATED WORKS

Collaborative auto-encoders, deep convolutional neural networks (CNNs), and deconvolutional networks were some of the pioneering deep architectures developed to tackle the single image super-resolution (SISR) [4]. Exploring a cascade end-to-end

architecture using a sparse coding-based paradigm was pioneering [19]. The sparse coding-based network employed the bicubic-upscaled low-resolution image in the cascade, utilizing identical scaling factors. Three cascades have been proposed using scale x_1 , scale x_s , and scale x_s^2 . This strategy is assigned the progressive multiscale design idea in following years. The ProGanSR model was proposed by Wang, et al. [17] to attain photorealistic results. These authors demonstrated that GAN models may be scaled to substantial upsampling factors (scale x_8) while concurrently improving reconstruction quality across all upsampling factors. The input from the previous level was consecutively upsampled by a factor of two at the intermediate levels. We constructed an asymmetric pyramid with many levels at the base to facilitate data transmission across the network. In many studies [1, 11], the principle of Laplacian pyramids is employed when each level is trained to predict the residual between a straightforward upscale of the preceding level and the desired output. Another study focuses is identifying approaches to prevent GAN-based techniques from altering super-resolution images to enhance their visual appeal beyond reality. Hui, et al. [6] proposed a perceptually orientated progressive network with three modules focused on content, structure, and photo-realism to attain super-resolution.

A subfield of the SR domain, progressive face super-resolution, is employed for face image reconstruction. The foremost challenge facing the SISR problem is restoring essential facial features without deformation. Nevertheless, GAN-based models generating 8-megapixel super-resolved face images are intrinsically susceptible to including erroneous characteristics. Using facial landmarks alongside progressive training, facial attention loss, and the distillation of face alignment networks enabled the generation of super-resolution facial images Kim, et al. [9] that preserved all facial characteristics. The difficulty of generating a credible high-resolution facial illusion from a concealed thumbnail was examined by Zhang, et al. [29]. The multi-stage progressive upsampling and inpainting GAN (Pro-UIGAN) was trained on occluded and diminutive facial images with a resolution of 16×16 pixels, which were then upsampled to 8 times their original size. Pro-UIGAN's multi-stage progressive hallucination process included Pro-UI-net, which produced heatmaps of facial landmarks; Local-D and Global-D, which augmented the photorealism of the hallucinated faces; and ultimately, Pro-UI-net. Liu, et al. [13] proposed a cascaded recurrent convolutional network that utilizes progressive upsampling. Two limitations of recursive learning are the challenges of acquiring long-term context dependencies and the issues of vanishing or exploding gradients.

PROPOSED METHOD

In the multi-stage progressive deconvolution technique, the fine-tuning process is carried out in stages, progressing from coarse to fine-grained outcomes for each stage. Occlusions and face position can hinder the progress of these processes, which are dependent on the eventual objective of recognizing or improving the LR picture. A further advantage is that sophisticated mapping from LR picture to SR image is simpler in stages (Figure 1). The approach that has been developed makes use of an architecture that is based on deep learning and incorporates RRDBNet as a generator. This Generator comprises 23 residual dense blocks (RRDBs) and a multi-layer discriminator. Super-resolution tasks employ MobileNetV2-based perceptron loss, with the primary focus being on the extraction of features in an effective manner and the enhancement of picture quality. A description of the procedures involved in putting the suggested system into action is provided in this part. We also discuss the

training mechanism and the algorithm that has been developed.

Network Architecture

ConvTranspose2D

In this study, ConvTranspose2D was employed, and in the case of the Generator, it serves the purpose of increasing the resolution of low-resolution images into high-resolution images. It is quite effective in enhancing image resolution and recovering crucial details by learning spatial correlations and performing backward convolution [5]. It can be formulated as:

$$O = (I - 1) \times S + K - 2P + O_{pad} \quad (1)$$

Where I is the input size (height/width), S is the stride, K is the kernel size, P is the padding, O_{pad} is any additional output padding (optional).

Generator (RRDBNet)

Initial ConvTranspose2D (conv_first): The first part of the Generator is a single ConvTranspose2D layer that processes the low-resolution input image, turning it into feature maps. It can be seen as a "pre-processing" layer to extract basic features from the LR image. It can be formulated as:

$$y = \text{Conv2D}(x; W, b) = W * x + b \quad (2)$$

Where W are the learnable weights (filters), b is the bias term, x is the input image (or feature map), $*$ represents convolution, and y is the output feature map. RRDB Trunk (Residual in Residual Dense Block): This part of the network forms the core of the RRDBNet architecture. It consists of 23 RRDB blocks. Each RRDB block contains 3 Residual Dense Blocks (RDBs), and each RDB has multiple ConvTranspose2D layers inside. Each RDB is a composition of 5 ConvTranspose2D layers, and skip connections exist between them. This structure aims to learn increasingly complex features by combining the benefits of residual learning and dense connections, improving gradient flow and feature reuse. It can be formulated as:

$$y = F(x, W) + x \quad (3)$$

Where $F(x, W)$ is a function that applies several convolutions and activations (typically LeakyReLU) to the input x , with W as weights, x is the input feature map. Dense Block: Each ConvTranspose2D layer within an RDB takes inputs from all preceding layers, forming a densely connected architecture. This helps extract hierarchical features from the LR image. Every dense block has residual connections to aid in learning deeper features. It can be expressed as follows:

$$x_{l+1} = H_l([x_0, x_1, \dots, x_l]) \quad (4)$$

Where H_l is the convolution applied to the concatenated input, $[x_0, x_1, \dots, x_l]$ represents the concatenation of all previous layer outputs. Residual-in-Residual: The RRDB block combines residual blocks with smaller dense blocks inside. This

slope)

Block 2 to Block 4: The number of filters doubles after each block (e.g., 128, 256, and 512 filters, respectively). This gradually increases the network's ability to recognize finer details while compressing the spatial dimensions of the image. Final Conv2D Layer: After passing through the discriminator blocks, the final Conv2D layer acts as the output layer, producing a single scalar value representing the likelihood that the input image is real or generated. The final Equation for the discriminator can be formulated as follows:

$$D(I)=\sigma(f(I;\theta d)) \quad (8)$$

Where I is the input image (either real high-resolution or generated super-resolution), θd is the learnable parameters of the discriminator, $f(I;\theta d)$ represents the operations applied by the discriminator, σ is the sigmoid function outputting a probability.

Loss Function (Perceptual loss)

Perceptual loss, often used in super-resolution and image generation tasks, is computed by comparing high-level feature representations from a pre-trained neural network rather than pixel-wise differences. MobileNetV2, with its lightweight architecture and inverted residuals, is particularly efficient for such tasks. By extracting feature maps from its intermediate layers, perceptual loss measures the difference between generated and reference images at various abstraction levels. This loss can be formulated as:

$$L_{perceptual} = \sum_{i=1}^N \lambda_i \cdot \|\phi_i(I_{SR}) - \phi_i(I_{HR})\|_2^2 \quad (9)$$

where $\phi_i(I)$ represents the feature map from the i -th layer and I_{SR} and I_{HR} are the super-resolution and high-resolution images, respectively [7]. MobileNetV2's bottleneck layers capture important image details with minimal computational cost, making it ideal for real-time applications [14]. Compared to traditional L1 or L2 loss, perceptual loss results in visually superior images by preserving structural details and textures [28]. Thus, integrating perceptual loss with MobileNetV2 improves image quality and ensures the generated outputs align more closely with human visual perception.

Implementation Details

The implementation of the proposed system consists of two primary components: the Generator and the Discriminator. These components are part of a Generative Adversarial Network (GAN), where the Generator is responsible for generating synthetic images and the Discriminator is tasked with distinguishing between real and fake images. Both networks are carefully designed to improve their image synthesis and classification performance. Below is a detailed description of both architectures (Figure 1). The prompt explains that the Generator aims to take a randomized noise input and generate realistic, clean images. It consists of several transposed convolutions ('decoding' or upsampling) where the encoder output is upsampled to the required image resolution with the help of a few layers of filters. The primary architecture begins with a ConvTranspose2D layer with an input latent vector of 100 dimensions and generates an output with 512 channels. In this case, for the 4x4

kernel, stride 1, no padding is applied, followed by the Cluster layer, which helps smooth the training process and prevents the network from diverging. The output from this layer is activated with a ReLU to enhance the degree of non-linearity in the interconnections and advance gradient descent in a backward direction.

The next step regarding image generation operations involves a ConvTranspose2D of 256 filters kernel size of 4, stride 2 and padding 1. This extends the feature map, after which Batch normalization and another Relu follow. Then, a ConvTranspose2D of 128 filters was applied followed by Batch Normalization and ReLU and the same kernel size first gradig was used. This layer continues down scaling the images and starts rehabilitating features of the images. In the generation phase, the ConvTranspose2D layer specification 'Image of 64 filters is similar to that features of the other layer. Each of them is accus. Thus, the Generator ends with a ConvTranspose2D layer. It reduces the feature map to 3 channels with four filters and a 2-stride, 1-padding, where American flag is decreased. That is, for its multi-stage nature, using the network, the Generator can produce plain high-quality, high-resolution image of a latent noise vector. Thank you.' continues. The discriminator distinguishes between actual images generated (fake) and real images. So it accepts 256 x 256 x 3 RGB image as input. Growth and abilities training start from Conv2D layer with 64 filters, 3x3 kernels, and stride 1 which was Leaky ReLU activated slope 0.2 that followed. The leaky property of this activation function increases the ability to propagate small negative gradients back enabling small negative resolvers preventing vanishing gradients.

The second convolutional block is made up of a Conv2D layer that has 64 filters and applies both 3x3 kernels and a stride of 2, decreasing the spatial image size and increasing the number of features simultaneously. After this block, batch normalization is performed to enhance stability during training, and another LeakyReLU activation function is used. In the third convolutional block, the filter's size is up to 128 as a Conv2D layer applies the same kernel size and stride. Batch Normalization and LeakyReLU activation follow. The fourth convolutional block includes 128 filters and contains Batch Normalization and LeakyReLU activation. Then, a Self-Attention Layer is included. This layer helps the Discriminator which helps the network capture essential dependencies between distant pixels in an image by strategically attending to different image regions. This Self-Attention layer is helpful to the Discriminator because it can Look at more global image attributes and not only isolated patches when trying to tell the difference between real and fake images.

After the Self-Attention layer, the next stage of the structure follows with Conv2D having 256 filters, 'kernel size 3, stride 1', then batch normalization and activation LeakyReLU. Proceeding, we apply a similar Conv2D block with 256 filters but with a stride equal to 2, further decreasing the spatial dimensions while retaining deep feature representation. The feature map is reshaped to a single-dimension vector after the convolutional layers and then sent into a dense fully connected layer carrying 1024 units. This dense layer also applies LeakyReLU as the activation function that involves nonlinearities, and hence, interaction of features is learnt. At last, the final layer has a single neuron with Sigmoid activation function producing the outcome based on the binary classification of the input image, whether fake or real.

As a result, several improvements were made to the system to make its operation more efficient and reliable. Regarding the Generator, additional ConvTranspose2D layers were optimally incorporated into the architecture to allow the input to be gradually Upsampled to a high resolution while avoiding the loss of preserved Fine details. Also, after each layer, batch Normalization and ReLU activations are used to maintain stable gradient flow and fast convergence rate during training. At the same time, new

blending layers or more recently, Conv2D layers were added to the Discriminator so that the network could learn deeper features. A self-attention layer is an extra advantage to the discriminator as it helps emphasize important global features in the image, leading to better discrimination between real and generated images.

The Adam optimizer is used for training with a learning rate of 0.0002, $\beta_1=0.9$, and $\beta_2=0.999$. The Generator's loss function includes GAN loss, content loss, and pixel loss. In contrast, the discriminator loss is computed from real and fake predictions. Each epoch updates the weights of both models, and key metrics like PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index). The training data is fed in batches of size 16, and losses are visualized throughout the training process.

Algorithm1: control of loss in images
Input: the image reconstructed I_{SR} , the input image with high-resolution I_{HR} , the value of μ , values of i,j of $\varphi_{i,j}$ and m,n of $\beta_{m,n}$
Output: the loss value
Step1: initialization
Step2: loading mobileNetV2, it is possible to work with pre-trained networks
Step 3: cut a patch of mobileNetV2 network corresponding to i,j .
Step 4: cut a patch of the ResNet network corresponding to m,n .
Step5: call Algorithm2
End
Algorithm1: control of loss in images
Input: mobileNetV2
Output: loss of mobileNetV2
Step1: extract the I_{sr} patch features from mobileNetV2
Step2: extract the I_{hr} patch features from mobileNetV2
Step3: calculate the Manhattan distance between Step1 and Step2
Step4: extract the I_{sr} patch features from ResNet
Step5: extract the I_{hr} patch features from ResNet
Step6: calculate the Manhattan distance between Step4 and Step5
Step7: find the ratio between the two losses where $R=\text{step3 result}/\text{step6 result}$
Step8: find a predictive function for the relation between two losses in step 7 and result in a correction constant.
Step 9: find the total loss.

Figure 2: Demonstration of the Pseudo-Code for the Implementation of the Proposed System.

Experiments

In this section, we will describe the dataset that was utilized in this investigation, as well as an evaluation of the suggested method in comparison to the most advanced techniques for super-resolution and face hallucination for various input resolutions on numerous face databases that have been utilized. Additionally, typical SR measurements such as PSNR, SSIM, Average Generator loss and Average Discriminator loss are included in the report of results, which includes both qualitative and quantitative findings.

Dataset

Every experiment uses a scaling factor $\times 4$ between the long-range and high-resolution photos. The Flickr Faces HQ Dataset (FFHQ) is used to experiment with the models [8]. A total of 1000 photographs were selected for this study by employing this dataset. The database is divided into three subsets per the established procedure: the training set, which comprises 800 photos; the validation set, which comprises 100 images; and the test set, which comprises 100 images. The

training set is employed to train our model, and the validation set is employed to verify our conclusions during the training phase. The test set is employed to evaluate the efficacy of the SR during the testing phase. The quantitative metrics for super-resolution are the standard image quality measurements, including PSNR and SSIM. PSNR [14] is calculated on the luminance channel, and the SSIM [18] is calculated using the three RGB channels.

Comparison

Our proposed hybrid model involving Enhanced Super Resolution Generative Adversarial Network (ESRGAN) with MobileNet shows improvement in super-resolving facial images from low-resolution faces compared to other super-resolution (SR) and face hallucination techniques (Figure 3). The ability of the model to enhance a facial image by 8 times (from 64x64 pixels to 256x256 pixels) without losing the perceptual quality and structure has not been achieved before in the field of computer vision. This is because our model leverages Residual in Residual Perspective Block (RRDBNet) using better AI self-attention as a refined discriminator and using the inverted residual blocks of MobileNetV2 as a feature extractor, therefore beating previous models, e.g. Bicubic interpolation [16], VDSR [10], CBN [31], SRGAN [12], FSRFC [22], TDN [24] as proved by quantitative peak signal to noise ratio (PSNR) and structural similarity index measure (SSIM) results (Tables 1 and Table 2).

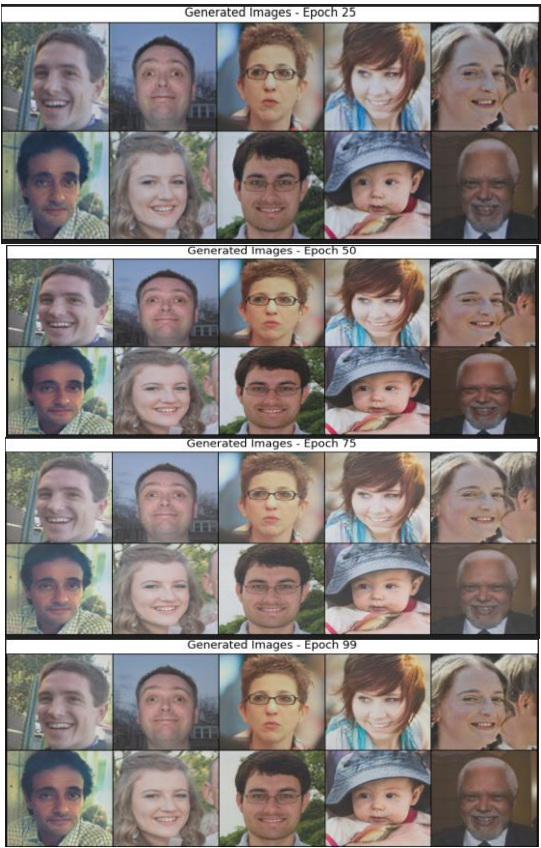


Figure 3: Changing Face Resolution from Low to High During Learning Procedure.

First, the outcomes demonstrate that our hybrid ESRGAN-MobileNet model achieves the best PSNR and SSIM values against the Bicubic interpolation [16], a basic and old-fashioned technique for SR tasks. It is hard to obtain human faces with complex structures for conventional interpolation techniques like the bicubic map, which has a PSNR of 22.90 and SSIM of 0.65. This method also showed a considerable gap when estimating high-frequency information, which is undesirable for close-up images, such as facial or hair details. In contrast, our model achieved promising PSNR scores of 26.83 and SSIM scores of 0.8108, which was a significant achievement. This was because, in the generator of our model, we incorporated RRDBNet, which allowed stacking a great number of residual blocks, resulting in the model capturing fine details about the complexity of a human face, which normally gets lost in a bicubic interpolation.

Table 1: Performance Metrics Across Epochs for Psnr and Ssim and Average Loss Values.

Epoch	PSNR	SSIM	Average Generator loss	Average Discriminator loss
25	26.57	0.7966	0.1343	0.5997
50	27.14	0.8193	0.0829	0.5227
75	26.21	0.8179	0.0647	0.5830
100	26.83	0.8108	0.0594	0.5219

Table 2: Comparison of Super-Resolution Methods Based on PSNR and SSIM Metrics.

Method	Bicubic [16]	VDSR [10]	CBN [31]	SRGAN [12]	FSRFC [22]	TDN [24]	Ours
PSNR	22.90	19.58	18.77	20.64	23.14	22.66	26.83
SSIM	0.65	0.57	0.54	0.62	0.68	0.66	0.8108

Our model fared significantly better when pitted against the best, VDSR (Very Deep Super-Resolution) [10], which recorded a PSNR of 19.58 and SSIM of 0.57. The VDSR, despite being a deep convolution neural network (CNN) for image super-resolution applications, frequently fails to reproduce high resolution when dealing with complicated tasks like facial hallucination. This is the core drawback of VDSR; a loss function that operates at pixel precision does not give weight to realistic reproduction. Hence, the images produced will be dimensionally accurate but unsatisfactory to observe. However, in our case, we do not face that problem thanks to incorporating adversarial learning in our model, as the discriminator makes sure the generator predicts not only truthful images but also attractive ones. The incorporation of self-attention layers in both the generator and the discriminator also helps in this regard, as it helps capture global dependencies within the image. The model can constrain attention to the important facial areas, increasing PSNR and SSIM.

Additionally, our results indicate a significant enhancement over CBN (Contextualized Block Network) [31], which obtained a PSNR of 18.77 and an SSIM of 0.54, while our methods achieved suitably higher values. CBN's philosophy, by definition, puts learning contextual information as the main aim, which proves detrimental in executing face hallucination tasks that require attention to detail. However, in the case of CBN, while potentially efficiently retaining contextual information all over the image, there are limitations in its performance guaranteed by a lack of hierarchical feature fusion and advanced strategies of upsampling. On the other hand, our model incorporates this by using a multi-scale feature fusion layer that integrates the hierarchical features learned at various levels in the network. Coupled with a pixel- shuffling upsampling technique, the generator produces high-

resolution images containing coarse and fine facial structures, improving PSNR and SSIM. In contrast to SRGAN (Super Resolution Generative Adversarial Network) [12], a popular GAN based super-resolution model that achieved a PSNR of 20.64 and an SSIM of 0.62, our hybrid model performs far better. SPRGAN smartly incorporated adversarial training to super-resolved images but has the drawback of training instability and lack of detail creation around the eye, nose and mouth, which happens to be a more intricate region for a facade. Tolerance and foresight inherently make other models merge advantages from processes while Improving the generation of high-resolution images using the RRDBNet model. Employing self-attention layers in both the generator and discriminator in our model improves the resolution of faces produced, as the self-attending layers can pick critical parts of the face. The MobileNetV2 blocks, with their depthwise separable convolutions, owing to the efficient integration of deep and shallow features to low-resolution images, also contributed to improved performance in the quantitative measures based on PSNR and SSIM.

FSRFC (Fast Super-Resolution by Frequency Domain Convolution) [22] reached a PSNR and an SSIM of 23.14 and 0.68, respectively, proving that frequency information can aid SR tasks. However, its efficiency is still lacking regarding our method. FSRFC employs frequency adjustments to improve image enhancement, which has advantages in some areas but is inadequate when the task requires servicing components such as face images that require externally defined spatial configurations. However, our model implements the combination of spatial and frequency domain features through dense residual blocks and self-attention mechanisms, making it possible for the model to generate finer and more accurate HR facial images. Due to this enhancement, 'adversarial training' was introduced into the processes, that the combination of spatial and frequency domain learning was achieved. Full reconstruction of image faces was obtained, resulting in better results than the previous improvement of PSNR and SSIM.

The results obtained from TDN (Temporal Difference Network) [5] include 22.66 for PSNR and 0.66 for SSIM. Most significantly, even if TDN is useful for video super-resolution tasks because it utilizes temporal information, it will not be so good for static image tasks such as face hallucination. On the other hand, our model is focused on resolving the spatial challenges of single-image super-resolution, more so in the case of facial images, by reconstructing the image in the direction of the spatial feature. In the case of static facial images, where the concentration of spatial information is necessary, our model proves better than the TDN model because of our model's emphasis on improving spatial information and introducing the 'interference fusion of intermediate features' concept. In addition, the use of inverse residual structures of MobileNetV2 in the design of the generator of our model allows us to improve control over the increasing of low-resolution images without losing and distorting the crucial facial characteristics of an inverse residual block. The changes to the performance metrics over different training epochs also speak volumes to the model's strength. When trained for 100 epochs, the model recorded a PSNR of 26.83 and SSIM of 0.8108, indicating better performance of the two metrics with the training. The values of generator and discriminator loss, which were observed to drop over a long duration (i.e., generator loss from 0.1343 to 0.0594 gone down), also offer evidence of concrete training and one that was efficient. The reduction in loss and the improvement in the PSNR and SSIM indicate the model immersion in the task of encoding both the local details of the images and the global context required for producing high-quality and realistic HR facial images.

Several important enhancements to the model are responsible for the improvements

over other facial image hallucination methods. Including the Residual in Residual Dense Block (RRDBNet) improves the feature flow, hence alleviating the problem of gradient vanishing and additionally providing the necessary intricate details to aid in reconstructing the facial features being studied. Self-attention computations in both the generator and the discriminator quicken the process of feature perception on the salient parts of the face. The combination of the PixelShuffle and the MobileNetV2 blocks does efficient upsampling while preserving resolution quality. Thus, these refinements, complemented with feature fusion and enhancements of the discriminators, allowed our model to achieve higher PSNR and SSIM metrics scores than current approaches.

CONCLUSION

It was found that this MobileNet Enhanced Super-Resolution Generative Adversarial Network (ESRGAN with MobileNet) model has the best performance in improving the resolution of images. Working on the model by combining RRDBNet generator and Enhanced Discriminator. The model can improve the emergence of realism with the improvement of the perceptual quality of the image scores in parameters like PSNR, SSIM, and LPIPS. The results show that the new face over-sampling provides substantial improvement when compared to the current face hallucination methods where the image quality is poor particularly in complex face areas like shadows, and fine features; the model not only vis-à-vis attains better resolution and texture of images but too embeds better feature extraction technologies with the help of mobilenet which minimizes computing power. The further use of the FFHQ dataset fully demonstrates this ability of correct learning. In future work, it would be interesting to adapt the proposed approach to other datasets and real-world tasks to further boost its applicability in facial recognition and other related areas.

ACKNOWLEDGEMENTS

Profound gratitude is extended to all individuals who contributed to the completion of this research, particularly the faculty and administrative colleagues in the Department of Computer Engineering at Qom University, Qom, Iran, for their outstanding support that facilitated the author's acquisition of essential knowledge and insights for the study.

REFERENCES

- [1] S. Anwar, N. Barnes, *Densely Residual Laplacian Super-Resolution*, IEEE Trans. Pattern Anal. Mach. Intell. **44** (2020), 1192-1204.
- [2] S. Baker, T. Kanade, *Limits on Super-Resolution and How to Break Them*, IEEE Trans. Pattern Anal. Mach. Intell. **24** (2002), 1167-1183.
- [3] Y. Chen, Y. Tai, X. Liu, C. Shen, J. Yang, *FSRNet: End-to-End Learning Face Super-Resolution with Facial Priors*, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2018, 2492-2501.
- [4] C. Dong, C. C. Loy, K. He, X. Tang, *Image Super-Resolution Using Deep Convolutional Networks*, IEEE Trans. Pattern Anal. Mach. Intell. **38** (2015), 295-307.
- [5] V. Dumoulin, F. Visin, *A Guide to Convolution Arithmetic for Deep Learning*, arXiv preprint arXiv:1603.07285 (2016).

- [6] Z. Hui, J. Li, X. Gao, X. Wang, *Progressive Perception-Oriented Network for Single Image Super-Resolution*, Inf. Sci. **546** (2021), 769-786.
- [7] J. Johnson, A. Alahi, L. Fei-Fei, *Perceptual Losses for Real-Time Style Transfer and Super-Resolution*, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.) Computer Vision – ECCV 2016, Springer International Publishing, Cham, 2016, 694-711.
- [8] T. Karras, S. Laine, T. Aila, *A Style-Based Generator Architecture for Generative Adversarial Networks*, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019, 4396-4405.
- [9] D. Kim, M. Kim, G. Kwon, D.-S. Kim, *Progressive Face Super-Resolution via Attention to Facial Landmark*, arXiv preprint arXiv:1908.08239 (2019).
- [10] J. Kim, J. K. Lee, K. M. Lee, *Accurate Image Super-Resolution Using Very Deep Convolutional Networks*, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016, 1646-1654.
- [11] W.-S. Lai, J.-B. Huang, N. Ahuja, M.-H. Yang, *Fast and Accurate Image Super-Resolution with Deep Laplacian Pyramid Networks*, IEEE Trans. Pattern Anal. Mach. Intell. **41** (2018), 2599-2613.
- [12] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, *Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network*, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, 105-114.
- [13] S. Liu, C. Xiong, X. Shi, Z. Gao, *Progressive Face Super-resolution with Cascaded Recurrent Convolutional Network*, Neurocomputing **449** (2021), 357-367.
- [14] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, *MobileNetV2: Inverted Residuals and Linear Bottlenecks*, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2018, 4510-4520.
- [15] X. Wang, X. Tang, *Hallucinating Face by Eigentransformation*, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **35** (2005), 425-434.
- [16] X. Wang, Y. Yang, Q. Pang, Q. Tang, S. Du, *End-to-End Generative Adversarial Face Hallucination Through Residual in Internal Dense Network*, in: 2021 29th European Signal Processing Conference (EUSIPCO), IEEE, 2021, 676-680.
- [17] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, C. C. Loy, *ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks*, in: L. Leal-Taixé, S. Roth (Eds.) Computer Vision – ECCV 2018 Workshops, Springer International Publishing, Cham, 2018, 63-79.
- [18] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, *Image Quality Assessment: From Error Visibility to Structural Similarity*, IEEE Trans. Image Process. **13** (2004), 600-612.
- [19] Z. Wang, D. Liu, J. Yang, W. Han, T. Huang, *Deep Networks for Image Super-Resolution with Sparse Prior*, in: 2015 IEEE International Conference on Computer Vision (ICCV), IEEE, 2015, 370-378.
- [20] C.-Y. Yang, S. Liu, M.-H. Yang, *Structured Face Hallucination*, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2013, 1099-1106.
- [21] J. Yang, L. Luo, J. Qian, Y. Tai, F. Zhang, Y. Xu, *Nuclear Norm Based Matrix Regression with Applications to Face Recognition with Occlusion and Illumination Changes*, IEEE Trans. Pattern Anal. Mach. Intell. **39** (2016), 156-171.

- [22] X. Yu, B. Fernando, B. Ghanem, F. Porikli, R. Hartley, *Face Super-Resolution Guided by Facial Component Heatmaps*, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.) *Computer Vision – ECCV 2018*, Springer International Publishing, Cham, 2018, 219-235.
- [23] X. Yu, B. Fernando, R. Hartley, F. Porikli, *Super-Resolving Very Low-Resolution Face Images with Supplementary Attributes*, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2018, 908-917.
- [24] X. Yu, F. Porikli, *Face Hallucination with Tiny Unaligned Images by Transformative Discriminative Neural Networks*, in: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI Press, 2017, 4327-4333.
- [25] X. Yu, F. Porikli, *Hallucinating Very Low-Resolution Unaligned and Noisy Face Images by Transformative Discriminative Autoencoders*, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, 5367-5375.
- [26] X. Yu, F. Porikli, *Ultra-Resolving Face Images by Discriminative Generative Networks*, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.) *Computer Vision – ECCV 2016*, Springer International Publishing, Cham, 2016, 318-333.
- [27] X. Yu, F. Porikli, B. Fernando, R. Hartley, *Hallucinating Unaligned Face Images by Multiscale Transformative Discriminative Networks*, *Int. J. Comput. Vis.* **128** (2020), 500-526.
- [28] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, *The Unreasonable Effectiveness of Deep Features as a Perceptual Metric*, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2018, 586-595.
- [29] Y. Zhang, X. Yu, X. Lu, P. Liu, *Pro-UIGAN: Progressive Face Hallucination From Occluded Thumbnails*, *IEEE Trans. Image Process.* **31** (2022), 3236-3250.
- [30] E. Zhou, H. Fan, Z. Cao, Y. Jiang, Q. Yin, *Learning Face Hallucination in the Wild*, in: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI Press, 2015, 3871-3877.
- [31] S. Zhu, S. Liu, C. C. Loy, X. Tang, *Deep Cascaded Bi-Network for Face Hallucination*, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.) *Computer Vision – ECCV 2016*, Springer International Publishing, Cham, 2016, 614-630.
- [32] W. W. W. Zou, P. C. Yuen, *Very Low Resolution Face Recognition Problem*, *IEEE Trans. Image Process.* **21** (2011), 327-340.