AGA-GAN: Attribute Guided Attention Generative Adversarial Network with U-Net for Face Hallucination

Abhishek Srivastava[a], Sukalpa Chanda[b], Umapada Pal[c]

[a]*Machine and Hybrid Intelligence Lab, Northwestern University, Chicago, USA*
[b]*Department of Computer Science and Communication, Østfold University College, Halden, Norway*
[c]*Computer Vision and Pattern Recognition Unit,Indian Statistical Institute, Kolkata, 700108, West Bengal, India*

Abstract

The performance of facial super-resolution methods relies on their ability to recover facial structures and salient features effectively. Even though the convolutional neural network and generative adversarial network-based methods deliver impressive performances on face hallucination tasks, the ability to use attributes associated with the low-resolution images to improve performance is unsatisfactory. In this paper, we propose an Attribute Guided Attention Generative Adversarial Network which employs novel attribute guided attention (AGA) modules to identify and focus the generation process on various facial features in the image. Stacking multiple AGA modules enables the recovery of both high and low-level facial structures. We design the discriminator to learn discriminative features by exploiting the relationship between the high-resolution image and their corresponding facial attribute annotations. We then explore the use of U-Net based architecture to refine existing predictions and synthesize further facial details. Extensive experiments across several metrics show that our AGA-GAN and AGA-GAN+U-Net framework outperforms several other cutting-edge face hallucination state-of-the-art methods. We also demonstrate the viability of our method when every attribute descriptor is not known and thus, establishing its application in real-world scenarios. Our code is available at https://github.com/NoviceMAn-prog/AGA-GAN.

Email addresses: abhishek.srivastava@northwestern.edu (Abhishek Srivastava), sukalpa@ieee.org (Sukalpa Chanda), umapada@isical.ac.in (Umapada Pal)

Keywords: Face hallucination, Generative adversarial network, U-Net, Spatial attention

1. Introduction

Face Hallucination is a domain-specific super-resolution task that aims to learn the mapping between low resolution (LR) image and its corresponding High resolution (HR) counterpart. It has great utility in a variety of applications like surveillance, face recognition, and expression recognition. Contrary to generic image super-resolution, the performance of face hallucination models relies heavily on their ability to effectively recover facial details and attributes. Maintaining the structural integrity of the face is imperative as the absence of it can lead to unnatural artifacts which may hamper the quality of the image. Particularly in surveillance, it has been empirically proven [1] that a minimum resolution ranging from 32 x 32 to 64 x 64 serves as a prerequisite for effective face recognition systems. In reality, video surveillance equipment may not be able to capture facial images satisfying the above criteria. This warrants a face hallucination architecture that can effectively upsample LR images.

In recent times this problem has drawn great attention [2]. It is worth mentioning here that while performing an 8x upsampling operation to produce an HR image, a method has only access to its LR counterpart which contains 1.56% pixels of the original image. While interpolation-based face hallucination techniques [3, 4] lacked the spatial details and visual quality, the presence of a large amount of HR and corresponding LR images enabled learning-based techniques [5, 6] to deliver promising results. The application of CNN [7, 8] and GAN [9, 10] based methods demonstrated superior performance with high visual quality and accurate structural representations. While usual methods exploit the prior structural information of the LR images in their generation process, the presence of attribute descriptors or labelled feature annotations can be further leveraged to enhance the visual quality of super-resolved images.

In this paper, we propose AGA-GAN(Attribute Guided Attention Generative Adversarial Network) which uses a novel attention mechanism to identify the spatial location of key facial attributes and focus the generation process to successfully recover salient facial elements. The AGA-GAN allows the attribute descriptors to learn their spatial position in the image domain

using the attribute stream. The attribute stream interacts with the main generation stream to allow feature maps carrying information about several facial elements to be incorporated in the final HR prediction. Attention maps generated by the attribute stream enable the main generation stream to focus on the spatial locations of essential features and also provide additional information about the description of facial elements. Feedback from the main generation stream allows the attribute stream to generate attention maps progressively focusing on various facial structures. Consequently, the super-resolved (SR) prediction possesses a high degree of fidelity and enjoys high-frequency details demonstrating superior visual quality. Apart from that, we design a spatial and channel attention-based U-Net [11] for enhancing the visual quality by refinement of facial features and even rectification of visual artifacts present in SR prediction by AGA-GAN (check Figure 9). It can be noted that apart from demonstrating a high degree of fidelity, the use of the U-Net module successively after AGA-GAN improves the quantitative performance (see Section 5.4). The contributions of our paper are summarized below:

1. We propose an attribute-guided attention-based technique that progressively refines the facial features in higher spatial dimensions to improve the quality of the generated SR image.
2. We propose a discriminator which leverages prior attribute descriptors to learn mappings between the descriptors and the HR/SR image to enhance its discriminative capacity.
3. We propose a U-Net-based architecture to take the prediction of our AGA-GAN and refine the image further as well as to increase the perceptual quality of the image.

The organization of the rest of the paper is as follows. Section 2 presents a brief literature survey of methods developed for face hallucination. Section 3 describes our proposed AGA-GAN+UNet framework. Our entire experiment setup is described in Section 4 and the results are reported in Section 5. Finally, the conclusion is presented in Section 6.

2. Related Work

2.1. Face Hallucination

Face hallucination has been explored over the past few decades and was originated by Baker and Kanade [12], where a multi-level learning model

based on the Gaussian image pyramid was used to increase the resolution of LR images. Liu et al. [13] used principal component analysis(PCA) and Markov random field(MRF) for face hallucination. Ensuring these pioneering works, various global face statistical methods [14, 15, 16] and local patch-based methods [6, 2] have been introduced for super-resolution of faces. Zhou et al. [17] initially used CNN to propose a bi-channel convolutional neural network (BCCNN) to hallucinate global face images. Chen et al. [18] proposed Residual back-projection network (RBPNet) which used a base model to extract features for face hallucination and edge map prediction boundaries. Huang et al. [19] exploited the fact that wavelet transform potentially represents the contextual and textural information of the image to design WaveletSRNet.They transformed the LR face images to wave coefficients and super-resolved the face image in the wavelet coefficient domain. Yu et al. [9] leveraged GAN [20] to develop URDGN which increased the perceptual quality of HR image. Indradi et al. [21] used inception residual networks inside the GAN framework to boost performance and stabilize training. Hi-FaceGAN [22] uses a suppression module for the selection of informative features which are then used by a replenishment module for recovery detail. SiGAN [10] uses two identical generators with pair-wise contrastive loss based on the fact that different LR face images possess different identities. Hence, they were able to super-resolve LR facial images while preserving their identities. Jiang et al. [23] used different deep learning-based approaches such as CNN, RNN, GAN to generate candidate SR images. Rather than using a pre-determined technique to combine the candidate predictions, ATMFN used an attention sub-network to learn the individual fusion weight matrices to determine useful components of candidate SR images. Further, a threshold-based fusion and reconstruction module combines the candidate HR image to give the final SR prediction. Chen et al. [24] proposed facial attention units(FAUs) which used a spatial attention mechanism to learn and focus on different face structures. Grm et al. [25] introduced C-SRIP (Cascade of simple Super-Resolution models) which used identity priors coupled with numerous super-resolution networks to progressively upscale the resolution. Since, C-SRIP gradually upscaled the resolution, it allowed identity constraints using pre-trained face recognition models at each step. Thereby, achieving better face hallucination results.

2.2. Attention Networks

Attention mechanisms are useful in identifying the most relevant features necessary for the effective completion of a task. Recently, this mechanism has been of great interest in the field of computer vision and has been studied extensively. Mnih et al. [26] devised a model capable of identifying a sequence of regions that convey the most relevant information. Hu et al. [27] proposed SE-Net which pioneered channel-wise attention. The Squeeze and Excitation (S&E) block was able to model inter-dependencies between the channels and derive a global information map that helps in emphasizing relevant features and suppressing irrelevant features. Xu et al. [28] devised a dual attention-based module for image captioning. Wang et al. [29] proposed to generate attention-aware features by stacking attention modules. Woo et al. [30] generated spatial and channel attention feature maps to multiply with input feature maps and refine them. Fu et al. [31] used an attention module that uses previous predictions as a reference while sequentially generating region attention. In the field of face hallucination, ATMFN [23] used an attention module to recognize the most relevant regions from all candidate SR images. SPARNet [24] used spatial attention in the generation process to progressively focus on various facial structures. Similarly, the DDA-SRGAN [32] combined channel and spatial attention mechanism to devise a dual-dimensional attention module which targeted discriminative features inside the region-of-interest(ROI). MA-SRGAN [33] employed an ROI-based loss function which required the generator to focus the generation process on creating visually pleasing features inside the pre-defined ROI.

3. Methodology

In this section, we present our proposed attribute-guided attention generative adversarial network. Here, LR, HR, SR, att, and s represent the low-resolution input, high-resolution target image, super-resolved prediction, attribute descriptors, and the upscale factor respectively. We explain various components of our two-stream generator. The generator aims to upscale the prior low resolution image LR (where $LR \in R^{W/s \times H/s \times 3}$) to high resolution space SR (where $SR \in R^{W \times H \times 3}$) with the assistance of att (where $att \in R^{38}$). The generator comprises a main stream and an attribute stream (Figure 1). The main stream takes the LR as input while the attribute stream takes the pair (LR and att) as input. The attribute stream subsequently helps

the main stream in generating an accurate high-resolution image with high fidelity. The attribute-guided attention module generates attention maps that guide the main stream to focus on the regions of facial attributes and maintains the structural integrity of the image. The feedback from the main stream to the attribute stream allows the attribute-guided attention maps to progressively create attention maps on various facial elements. Stacking these modules allows an increased receptive field of the attention maps enabling the generation stream to focus on both high and low-level facial structures. The residual structure of the module leads to faster convergence and improved information flow. We also propose a discriminator which uses att as a prior and tries to distinguish between all (HR, att) and (SR, att) pairs. The super-resolved image from the AGA-GAN along with bicubic interpolation of the LR image is concatenated and is then fed into spatial and channel attention U-Net. The U-Net aims to refine existing facial structures and add further high-frequency details to raise the perceptual quality and richness of generated SR predictions. We expand on our proposed approach in the subsequent sections.

3.1. Generator

The generator consists of two streams denoted as the main stream and attribute stream (see Figure 1). The main stream takes an LR image as input and the attribute stream takes an attribute vector as well as the corresponding LR image. The attribute vector provides information about the presence and absence of various facial features in the image. The main stream initially has 3 convolutional and ReLU layers. We then use residual dense blocks proposed in [34] to increase the effective receptive fields and extract high-level features from the input image while preserving low-level features. The extracted features are scaled by factor and added back to the input to avoid instability [35, 36]. The residual design of the blocks allows relevant high and low-level features to be preserved for effective up-sampling. The attribute stream translates the one-dimensional vector to two-dimensional image space and forces interaction with the LR image stream embedded in the attribute stream (Figure 1), thus enabling the features to capture the relationship between the facial attribute descriptors and the image. We then use sub-pixel convolution [8] to increase the spatial dimension by a factor of 2 simultaneously in the main and the attribute stream. Further, the attribute-guided attention module uses the attribute stream's feature maps to generate attention coefficients to learn the spatial correlation between attributes and

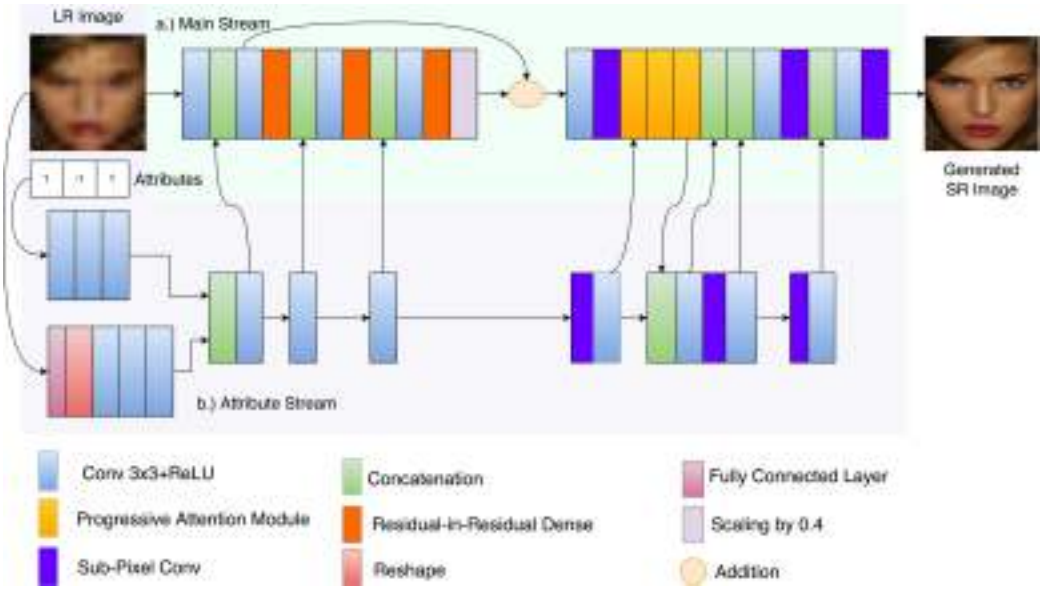

Figure 1: The Generator Architecture of our proposed Attribute Guided Attention Generative Adversarial Network. The b.) Attribute stream (lower stem) takes the LR image and the corresponding attribute feature descriptors as input. The a.) Main stream (upper stem) takes the LR image as input and leverages the attribute guided attention maps generated by the Attribute stream to refine facial structures and progressively produce high frequency features.

their corresponding location in the image domain. This enables the main stream to improve the facial and textural features of the SR image.

3.1.1. Attribute Guided Attention Module

In this section, we describe our attribute guided attention module (see Figure 2). The attribute stream utilizes the attribute vector and the LR image to learn the mapping between the features and their region of interest in the image domain. Initially, two stacked convolutional layers and a single convolutional layer are used to operate on sets of feature maps extracted from the main and attribute stream respectively, as explained in Equation 1 and Equation 2

$$SR_{main} = Conv(Conv(SR_{main})) \tag{1}$$

$$AS_{stem} = Conv(AS_{stem}) \tag{2}$$

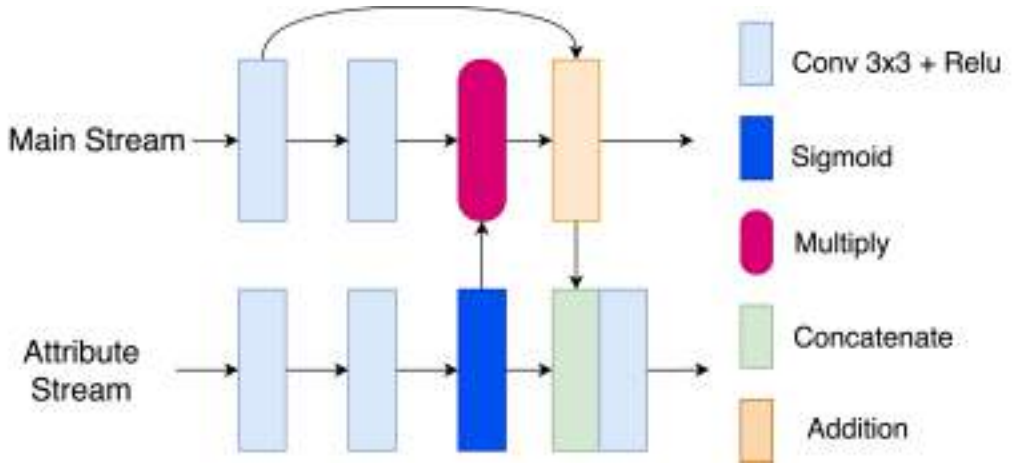

Figure 2: The Attribute Guided Attention Module, which generates the attribute guided attention(AGA) maps. The AGA maps are used to identify the spatial locations of the facial attributes during the generation process of the Main stream and provides assistance in progressively refining the facial structures.

We then, use the sigmoid activation function to calculate an attention coefficient for each spatial location in the feature maps as described in Equation 3.

$$AGA = \sigma(Conv_{1x1}(AS_{stem})) \tag{3}$$

The AGA map is multiplied with SR_{main} to identify the regions in the image domain corresponding to the facial attributes. This is termed as attribute enhanced feature maps in Equation 4 which is then added back to the SR_{main} (Equation 5). This residual connection preserves the relevant low-level structural information and improves convergence.

$$AEF = AGA \otimes SR_{main} \tag{4}$$

$$SR_{main} = SR_{main} + AEF \tag{5}$$

The refined SR_{main} is then fused with AS_{stem} to provide feedback to the attribute stream in Equation 6. This allows subsequent attribute-guided attention modules to progressively generate attention maps focusing on auxiliary facial elements. Consecutive attribute guided attention modules also provide an additional advantage in identifying high-level target structures by increasing the receptive field. We use three such modules in a sequence.

$$AS_{stem} = AS_{stem} \oplus SR_{main} \tag{6}$$

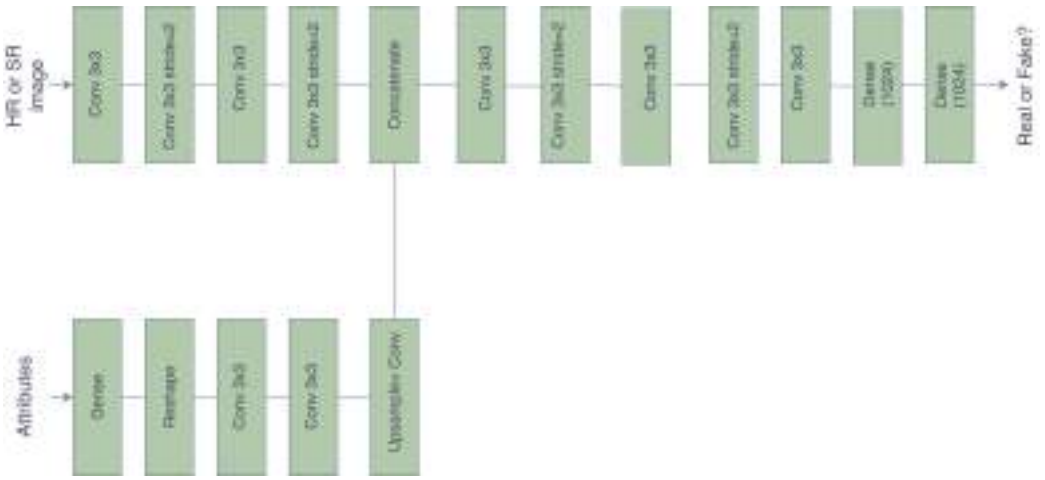

Figure 3: The Architecture of the Discriminator used in our AGA-GAN. The attribute descriptors (lower stem) is concatenated with the features extracted from the HR/SR image (upper stem) to further guide the discriminator to differentiate between HR and SR images.

3.1.2. De-convolution Sub-Network

The resultant feature maps by attribute guided attention modules of both main stream (SR_{main}) and attribute stream(AS_{stem}) are used for upscaling. The SR_{main} and AS_{stem} are both upsampled using sub-pixel convolution [8]. The main stream feature maps are then concatenated with attribute stream feature maps as described in Equation 7, this helps to further leverage the attribute information propagated in the upsampling process (refer Figure 1). We upsample the feature maps twice in the case of 8x upsampling and once in the case of 4x upsampling.

$$SR_{main} = SR_{main} \oplus AS_{stem} \tag{7}$$

3.2. Discriminator

We design our discriminator to take both the HR/SR image and the attribute descriptors of the corresponding image as input. This enables the discriminator to learn the correlation between the facial features and structures of the HR image and its corresponding attribute descriptors. The attribute prior then enables the discriminator to distinguish between the real HR image and the fake SR image based on that facial feature's existence

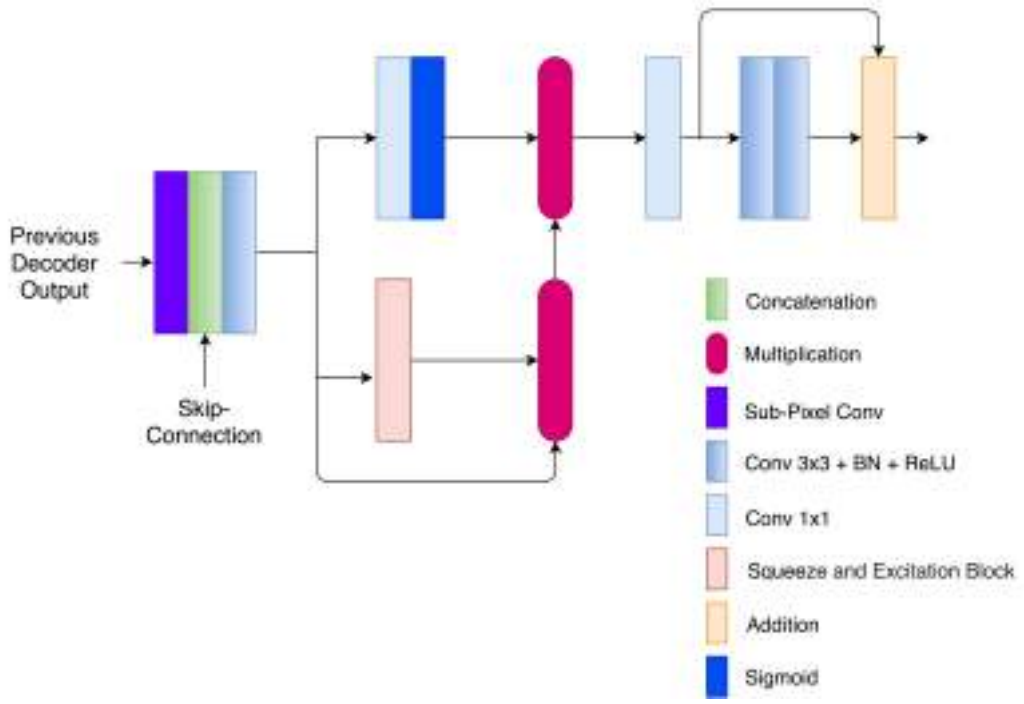

Figure 4: Structure of the Dual Attention Block used in our proposed method.

in the image domain as well as its shape and composition. There are again two branches that process the image and the attribute vector separately (see Figure 3). The attribute vector is first connected to a fully connected layer and is then reshaped into 16x16x3. The HR/SR image is fed into the main branch which comprises convolutional units. Each unit consists of a convolutional layer with a 3x3 kernel with LeakyReLU activation. This is followed by a convolutional layer with a kernel size of 3 and stride of 2 and a LeakyReLU activation function. When the image spatial dimensions are reduced from 128x128 to 32x32, the tensors from the main branch and attribute branch are concatenated to combine the information from the image domain and its attribute descriptors. This enables the relationship between the attribute priors and their presence, structure, and composition in the corresponding image to serve as discriminative features in distinguishing the real and the generated image.

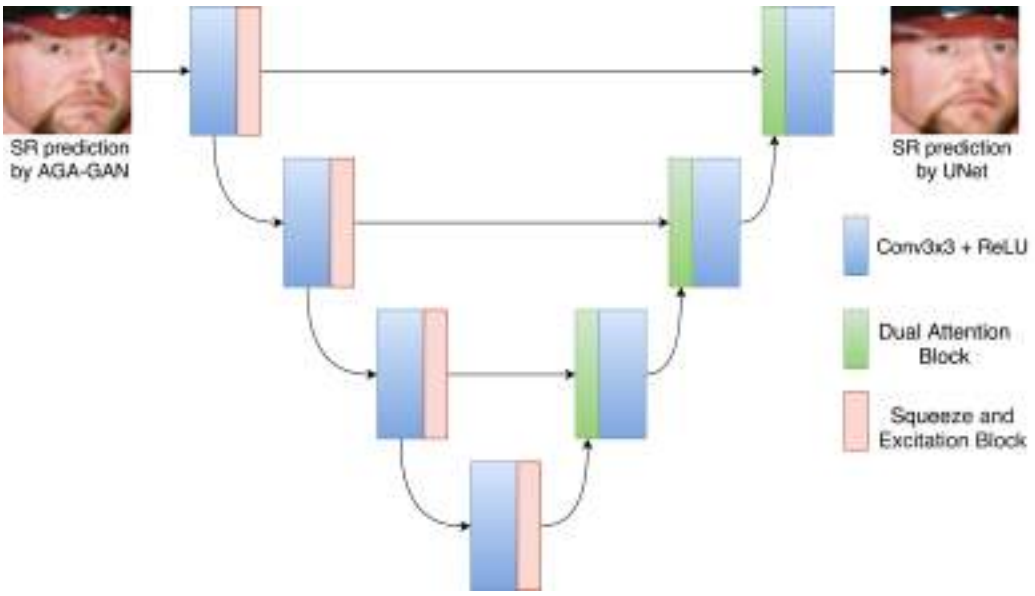

Figure 5: The Dual attention U-Net Architecture. At each level of the encoder, the squeeze and excitation module is used to apply channel-wise attention. At each level of the decoder, the dual attention block is used to prune irrelevant features and amplify the relevant features.

3.3. U-Net

We build a U-Net architecture (see Figure 5) which takes the SR prediction of AGA-GAN and bicubic interpolation of the LR image as input and uses the channel and spatial attention in the encoder-decoder architecture. The bicubic interpolation of the LR image is combined with the SR image by concatenation along the channel axis. We use an encoder-decoder-based network to refine the pre-existing SR image. The encoder block consists of convolutional layers followed by LeakyReLU activation and (S&E) layer. The (S&E) block in the network increases the network's representative power by computing the interdependencies between channels. During the squeezing step, global average pooling is used to aggregate feature maps across the channel's spatial dimensions. In the excitation step, a collection of per-channel weights are produced to capture channel-wise dependencies [27]. At each encoder stage, max pooling with the stride of 2 is used for downscaling the resolution, and drop out is used for the model regularization. The skip connections propagate low-level features from the encoder block to its cor-

responding decoder block from the same level. We also use dual attention blocks (see Figure 4) in the decoder level to provide spatial and channel attention to the features transmitted by the previous decoder blocks. The first attention mechanism applies channel attention, whereas the second attention uses a spatial attention mechanism. We have used a S&E block for the calculation of channel-wise scale coefficients denoted by X_{se}. Spatial attention is also calculated at the same top stream where the input channels C are reduced to 1 using 1×1 convolution. The sigmoid activation functions $\sigma(\cdot)$ is used to scale the values between 0 and 1 to produce an activation map, which is stacked C times to give X_c. The output of the spatial and channel attention can be represented as:

$$D_{sc} = (X_c + 1) \otimes X_{se}, \tag{8}$$

where $\otimes$ denotes the Hadamard product and X_c is increased by a magnitude of 1 to amplify relevant features determined by the activation map. D_{sc} represents the output of the dual attention block which is then fed into the decoder block which consists of convolutional layers of 3x3 kernel size followed by LeakyReLU activation and sub-pixel convolutional layers to increase the spatial dimension by a factor of 2. The output of the final decoder block is processed by a 1x1 convolutional layer which decreases the output number of channels to 3 and applies tanh activation. We find that this improves the performance of the network even further and increases the fidelity and visual quality of the generated images (see Section 5.4).

3.4. Loss Function

We denote the SR image prediction by AGA-GAN as SR_{AGA} and the SR prediction by U-Net as SR_U. The initial discriminator output is $D(x)$ which is the probability that the input image x is real. The loss for the discriminator is given as:

$$L_D = \sum_{n=1}^{N} [-\log(D(HR) - \log(1 - D(SR_{AGA}))] \tag{9}$$

The loss for the generator is given as:

$$L_G = \sum_{n=1}^{N} \log(1 - D(SR_{AGA})) \tag{10}$$

where, N represents the total data samples. Pixel-wise mean absolute error (MAE) loss is calculated as:

$$L_{mae}(P,Q) = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |P_{x,y} - Q_{x,y}| \tag{11}$$

We use VGG-19 [37] pre-trained on ImageNet [38] to gather feature representations of the input image. The Euclidean distance between the feature representations of the HR image and the SR image serves as the perceptual loss L_{percep}. The total loss function of the generator is defined as:

$$\begin{aligned} L_{total} = 0.003 * L_{GAN} + L_{percep}(HR, SR_{AGA}) \\ + L_{mae}(HR, SR_{AGA}) \end{aligned} \tag{12}$$

Similarly the loss in the U-Net stage is,

$$L_{U-Net} = L_{percep}(HR, SR_U) + L_{mse}(HR, SR_U) \tag{13}$$

4. Experiments

4.1. Implementation Details

We use one-sided label smoothing [39] to encourage the discriminator to predict soft probabilities. Preventing the discriminator from being overconfident in its prediction stabilizes the training and also provides regularization [40]. MTCNN [41] is used for face alignment as the initial pre-processing step. We resize the images to 128 x 128 and downsample them to 16 x 16 and 32 x 32 for 8x and 4x upscaling tasks respectively, before feeding it into the model where the batch size used is 50. The images are normalized to a range of [-1,1] and tanh activation is used to generate the final prediction which maps the image in the range of [-1,1]. The output of residual in residual dense blocks is scaled by a factor of 0.4 before adding it into the input. The Adam optimizer was used with a constant learning rate of $1e^{-4}$, and a dropout regularization with $p = 0.2$ was used. All models are trained for 50 epochs. The proposed model is available at https://github.com/NoviceMAnprog/AGA-GAN.

4.2. Evaluation metrics

Standard computer vision metrics for Face hallucination such as Peak signal-to-noise ratio (PSNR), Structural Similarity Index (SSIM) have been used for our experiments. We have also used Feature-based similarity index (FSIM) [42], Signal to reconstruction error ratio (SRE) [43], Universal image quality index (UIQ) [44] for a detailed and thorough comparison with other SOTA methods. Additionally, we have provided BRISQUE score for all methods, which is a no-reference metric for evaluating image quality. The BRISQUE score of an image is inversely proportional to the image quality [45]. For evaluating the performance of the hallucination models on downstream tasks, we use FaceNet-512 [46] to perform face verification. Subsequent cosine distance is calculated and reported between the ground truth HR and the generated SR images.

4.3. Dataset

We use CelebA dataset [47] for our experiments. The CelebA dataset consists of 202,599 images and each image is accompanied by 40 binary attribute annotations. The images carry large fluctuations in pose and background clutter. For our experiments, we choose 100,000 images for training and 10,000 images for testing. Each image has 38 attributes that describe the features of the image as 2 classes fall out of the region after preprocessing. For attribute-specific evaluation, 1000 images that have "big nose", 500 images with "eyeglasses", 400 images that possess a "goatee", 500 images with a "mustache", and 1000 images that have "narrow eyes" are selected for the experiments.

5. Results and Discussion

To demonstrate the effectiveness of our proposed method we have conducted several experiments which are structured as follows. We report the performance of our method in 8x upscaling and 4x upscaling tasks in Subsection 5.1.1 and Sub-section 5.1.2, respectively to exhibit the superior performance of AGA-GAN and AGA-GAN+U-Net in face hallucination tasks in comparison to other published methods. We also evaluate the performance of our method on attribute-specific images in Sub-Section 5.2 to illustrate the network's capacity to recover challenging facial attributes. Evaluation and comparison of AGA-GAN+U-Net against other baselines on downstream face verification task is reported in Sub-section 5.3. We compare the advantages

of using U-Net with AGA-GAN for face hallucination in Sub-section 5.4. In Sub-section 5.5, we analyze the performance of our proposed method when only partial attributes are present. To show further practicality of our method in real-world scenarios, we study the performance of AGA-GAN when partial attributes are known in Sub-section 5.5.

5.1. Comparison with Other Published Methods

In this section, we describe the methods that are used for the quantitative and qualitative comparison with AGA-GAN and AGA-GAN+U-Net.

- AACNN [48] also uses the attributes and their interaction with LR image for the generation of SR predictions.
- SPARNet [24], which uses a spatial attention mechanism to focus the generation process on key face structure regions.
- DDA-SRGAN [32] uses a dual dimensional attention mechanism to combine channel and spatial attention and focus the generation process on the more discriminative features.
- MA-SRGAN [33] combines the nESRGAN+ [49] with an ROI based loss function.
- ATMFN [23] which unified CNN, RNN, and GAN-based super-resolvers and an attention sub network to gather the most informative regions. The combination of these candidates was formed using threshold-based fusion to give the final HR prediction.
- SiGAN [10] uses dual generators and discriminators which ensures the preservation of identities while making the HR prediction.

We use the author-released codes of these methods for our experiments. Further, these models are trained under the same settings and follow the same train-test split to provide a fair comparison. In all the Tables reporting the quantitative comparison of our method with other state-of-the-art methods, the best performance is in bold and second-best performance is highlighted in blue.

5.1.1. Comparison with the scale factor of 8

In this experiment we aim to upscale the image by a factor of 8, the original 128 x 128 images are downsampled by a factor of 8 before feeding it into our AGA-GAN+U-Net framework. From quantitative results in Table 1, we can observe that AGA-GAN and AGA-GAN+U-Net outperform all other SOTA methods across all calculated metrics. The qualitative comparison with other methods on the CelebA dataset is shown in Figure 6. We can observe that bicubic interpolation of the image causes a major loss in facial details of the face whereas, SparNet with the help of spatial attention mechanism generates finer facial features. While DDA-SRGAN is able to capture the structural features of the face, it fails to recover high-frequency features and therefore generates SR images with aberrations and artifacts. Similarly, ATMFN effectively captures the facial structure but lacks detailed features causing the image to be blurry. AACNN and MA-SRGAN can generate detailed and fine images but some artifacts persist. Overall our AGA-GAN and AGA-GAN+U-Net produce sharper images of superior visual quality, devoid of disturbing artifacts. It can be observed from Table 1 that AGA-GAN and AGA-GAN+UNet has only 9.77M and 13.63M parameters which is significantly lower than SparNet [24], ATMFN [23] and comparable to SiGAN [10].

Table 1: Result comparison on the Celeb-A dataset with an upscale ratio of 8.

Method	**PSNR**	**SSIM**	**FSIM**	**SRE**	**UIQ**	**BRISQUE**	**Parameters**
Bicubic	29.2394	0.5568	0.5137	49.2655	0.3573	89.6799	-
SiGAN [10]	28.9507	0.6364	0.5495	49.0489	0.4653	42.2978	13.38M
AACNN [48]	30.2314	0.6819	0.5863	51.2364	0.5378	38.2990	2.49M
DDA-SRGAN [32]	29.5629	0.6026	0.5317	49.7240	0.4260	39.1042	1.10M
MA-SRGAN [33]	30.6219	0.7304	0.6108	51.7094	0.5891	41.3805	3.95M
SparNet [24]	29.2508	0.6692	0.5841	50.2564	0.5206	43.1848	25.44M
ATMFN [23]	29.6547	0.6367	0.5492	48.8089	0.4727	80.7350	64.26M
AGA-GAN (Ours)	31.4405	0.7848	0.6509	52.4967	0.6679	37.0216	9.77M
AGA-GAN+U-Net(Ours)	**31.9171**	**0.8149**	**0.6670**	**52.8662**	**0.7079**	**31.2508**	13.63M

5.1.2. Comparison with the scale factor of 4

In this experiment, we aim to upscale the image by a factor of 4, the original 128 x 128 images are downsampled by a factor of 4 to 32 x 32 before it is coupled with its corresponding attribute vector and served as input to our model. The quantitative results are provided in Table 2 and the qualitative results are shown in Figure 7. We can observe from Table 2 that

Figure 6: Qualitative Comparison of AGA-GAN with other SOTA methods on 8x upscaling.

AGA-GAN+U-Net surpasses all other SOTA methods in all metrics except BRISQUE. In Figure 7 we can observe that each method is effective to recover face structures to an extent. AACNN and MA-SRGAN can generate fine details but noise and irregularities in the facial structures are observed. However, it must be noted that MA-SRGAN was able to achieve the best BRISQUE score for 4x upscaling. SiGAN and ATMFN synthesize blurry and fuzzy predictions. Although SPARNet demonstrates decent capacity in recovering sharper features which can be attributed to its facial attention units, the predictions are susceptible to noise. The super-resolved images by DDA-SRGAN achieves a better PSNR and UIQ as compared to AGA-GAN for 4x upscaling (see Table 2). Bicubic interpolation lacks the desired feature details in its predictions. AGA-GAN+U-Net exhibits superior performance by generating high-quality SR predictions. The images accurately capture the facial structures and are enriched with fine and intricate details. Consequently, the images possess a high degree of fidelity.

5.2. Evaluation on Attribute Specific Facial Images

In this section, we study the performance of our selected methods against our AGA-GAN+U-Net framework on subsets of the Celeb-A dataset which

Figure 7: Qualitative Comparison of AGA-GAN with other SOTA methods on 4x upscaling.

Table 2: Result comparison on the Celeb-A dataset with an upscale ratio of 4.

Method	**PSNR**	**SSIM**	**FSIM**	**SRE**	**UIQ**	**BRISQUE**	**Parameters**
Bicubic	30.6845	0.7513	0.6315	51.2223	0.6463	60.2453	-
SiGAN [10]	29.1226	0.5881	0.5749	50.5577	0.4783	39.5480	3.27M
AACNN [48]	32.4497	0.8703	0.7104	53.6478	0.7892	43.8673	2.34M
DDA-SRGAN [32]	34.4866	0.9062	0.7517	53.9451	0.8641	51.3549	1.09M
MA-SRGAN [33]	32.3944	0.8465	0.7181	53.7677	0.7653	**28.6274**	3.80M
SparNet [24]	30.3309	0.7584	0.6382	51.7745	0.6497	40.6698	25.44M
ATMFN [23]	30.7626	0.7624	0.6197	49.7543	0.6715	70.2202	78.21M
AGA-GAN (Ours)	34.1323	0.9104	0.7522	54.9578	0.8630	33.7918	9.26M
AGA-GAN+U-Net(Ours)	**34.7659**	**0.9247**	**0.7676**	**55.3552**	**0.8848**	36.0212	13.12M

contains specific attributes. We present both quantitative and qualitative comparisons of each method's competency in recovering facial attributes from the input LR image. We ensure that these subsets maintain maximum overlap with the original test set on the 8x upscaling task. The detailed qualitative comparison is presented in Figure 8. We choose the attributes "big nose", "eyeglasses", "goatee", "slightly open mouth", "mustache" and "narrow eyes" as our primary facial attributes to be recovered in this experiment. These attributes cover the eye, nose, and mouth region of the face and

Table 3: Result comparison on facial images with "big nose" attribute.

Method	PSNR	SSIM	FSIM	SRE	UIQ	BRISQUE	Parameters
Bicubic	29.1823	0.5314	0.5027	48.6661	0.3353	89.5779	-
SiGAN [10]	28.8972	0.6067	0.5353	48.3891	0.4422	42.2193	13.38M
AACNN [48]	30.0141	0.6450	0.5676	50.4107	0.5078	36.6382	2.49M
DDA-SRGAN [32]	29.4561	0.5715	0.5173	49.0186	0.4021	37.0030	1.10M
MASRGAN [33]	30.4044	0.7002	0.5942	50.9324	0.5635	42.5033	3.95M
SparNet [24]	29.2007	0.6407	0.5690	49.5539	0.5010	43.7745	25.44M
ATMFN [23]	29.7448	0.6163	0.5411	48.2906	0.4634	80.7797	64.26M
AGA-GAN(Ours)	31.0675	0.7519	0.6309	51.6057	0.6378	36.1596	9.77M
AGA-GAN+U-Net(Ours)	**31.5187**	**0.7848**	**0.6446**	**51.9691**	**0.6777**	**30.3936**	13.63M

possess a varying amount of difficulty in recovery. In Table 3 we report the quantitative comparison of AGA-GAN+U-Net framework with other methods on images that have "big nose". We can observe that the performance of all methods drops as compared to the global 8x upscaling experiment(see Table 1) except for ATMFN whose performance improves. AGA-GAN+U-Net reports the best performance across all metrics in this category. In Figure 8 we can observe that each method can recover and determine the size of the nose to an extent. We notice that AGA-GAN and AGA-GAN+U-Net are effective in recovering structures and detailed features of the target attribute.

Table 5 and Table 4 report the performance of each method on images that possess eyeglasses and have narrow eyes as facial features, respectively. We can observe that these attributes are relatively hard to recover with faces having eyeglasses reporting the lowest results across all metrics and methods as compared to other attributes. This is also supported by Figure 8 where we can see that most methods are unable to generate the eyeglasses. This can be attributed to the fact that the LR image is 16 x 16, which may contain a very limited amount of information regarding eyeglasses. Even though attribute descriptors relay the information about the presence of eyeglasses to AACNN's network, it still fails to retrieve them. AGA-GAN and AGA-GAN+U-Net, leveraging the attribute-guided attention module is effective in recovering this attribute. While AGA-GAN lacks the finer details, AGA-GAN+U-Net can further refine and generate sharper and visually pleasing details of the eyeglasses. Another region we explore is the mouth area of the facial images. Attributes that describe the facial hair of the image i.e. "goatee" and "mustache" are hardest to recover in this region as evident in

Figure 8: Qualitative Comparison of AGA-GAN+U-Net with other SOTA methods on specific facial attributes.

Table 4: Result comparison on facial images with "narrow eyes" attribute.

Method	**PSNR**	**SSIM**	**FSIM**	**SRE**	**UIQ**	**BRISQUE**	**Parameters**
Bicubic	29.2066	0.5480	0.5131	49.2006	0.3500	89.3715	-
SiGAN [10]	28.9139	0.6311	0.5500	49.0028	0.4619	42.7260	13.38M
AACNN [48]	30.1428	0.6732	0.5828	51.1525	0.5306	38.7267	2.49M
DDA-SRGAN [32]	29.5133	0.5988	0.5322	49.7315	0.4231	50.4031	1.10M
MASRGAN [33]	30.5381	0.7258	0.6097	51.6448	0.5859	43.3568	3.95M
SparNet [24]	29.2244	0.6623	0.5821	50.1950	0.5156	34.6491	25.44M
ATMFN [23]	29.6100	0.6301	0.5478	48.7321	0.4676	80.1130	64.26M
AGA-GAN (Ours)	31.3206	0.7790	0.6482	52.4059	0.6634	38.8545	9.77M
AGA-GAN+U-Net(Ours)	**31.7804**	**0.8090**	**0.6642**	**52.7705**	**0.7028**	**32.2972**	13.63M

Table 6 and Table 8 whereas, "slightly open mouth" is effectively recovered as shown in Table 7. In Figure 8 we can note that even though these attributes are prominent in each method's predictions, they lack the finer details present in the original HR image.

Table 5: Result comparison on facial images with "eyeglasses" attribute.

Method	PSNR	SSIM	FSIM	SRE	UIQ	BRISQUE	Parameters
Bicubic	29.1504	0.5072	0.4947	48.2397	0.3146	89.2502	-
SiGAN [10]	28.8148	0.5818	0.5288	48.1265	0.4108	43.8321	13.38M
AACNN [48]	29.8511	0.6096	0.5547	49.9227	0.4662	39.9565	2.49M
DDA-SRGAN [32]	29.2901	0.5312	0.5021	48.3467	0.3671	33.4407	1.10M
MASRGAN [33]	30.1717	0.6651	0.5791	50.3789	0.5219	46.1387	3.95M
SparNet [24]	29.1150	0.5967	0.5520	49.0468	0.4537	34.9088	25.44M
ATMFN [23]	29.4592	0.5755	0.5203	47.7411	0.4184	78.0582	64.26M
AGA-GAN (Ours)	30.7578	0.7122	0.6104	50.9109	0.5882	36.9800	9.77M
AGA-GAN+U-Net(Ours)	**31.1469**	**0.7468**	**0.6245**	**51.2552**	**0.6298**	**30.9812**	13.63M

Table 6: Result comparison on facial images with "goatee" attribute.

Method	PSNR	SSIM	FSIM	SRE	UIQ	BRISQUE	Parameters
Bicubic	29.1275	0.5209	0.4989	48.1002	0.3360	89.5204	-
SiGAN [10]	28.8812	0.5942	0.5302	47.9543	0.4354	42.6179	13.38M
AACNN [48]	29.8320	0.6151	0.5512	49.7123	0.4840	32.3641	2.49M
DDA-SRGAN [32]	29.3693	0.5558	0.5102	48.4121	0.3957	34.4117	1.10M
MASRGAN [33]	30.2535	0.6834	0.5845	50.3450	0.5529	43.9116	3.95M
SparNet [24]	29.1498	0.6207	0.5570	48.9490	0.4875	33.8498	25.44M
ATMFN [23]	29.4613	0.5976	0.5311	47.7364	0.4490	81.2881	64.26M
AGA-GAN (Ours)	30.8354	0.7333	0.61758	50.9577	0.6232	37.0496	9.77M
AGA-GAN+U-Net(Ours)	**31.2479**	**0.7663**	**0.6309**	**51.3081**	**0.6625**	**31.0462**	13.63M

Table 7: Result comparison on facial images with "slightly open mouth" attribute.

Method	PSNR	SSIM	FSIM	SRE	UIQ	BRISQUE	Parameters
Bicubic	29.1911	0.5462	0.5108	49.2990	0.3504	89.4186	-
SiGAN [10]	28.9329	0.6276	0.5471	49.0956	0.4620	42.1772	13.38M
AACNN [48]	30.1581	0.6744	0.5839	51.2004	0.5365	37.8457	2.49M
DDA-SRGAN [32]	29.5168	0.5929	0.5295	49.7965	0.4192	35.5552	1.10M
MASRGAN [33]	30.5299	0.7213	0.6070	51.6356	0.5855	40.7270	3.95M
SparNet [24]	29.2162	0.6650	0.5834	50.2908	0.5224	33.9014	25.44M
ATMFN [23]	29.5945	0.6299	0.5493	48.9114	0.4691	80.1840	64.26M
AGA-GAN (Ours)	31.3236	0.7799	0.6492	52.4407	0.6685	36.3803	9.77M
AGA-GAN+U-Net(Ours)	**31.7866**	**0.8102**	**0.6653**	**52.8103**	**0.7081**	**30.1729**	13.63M

Table 8: Result comparison on facial images "mustache" attribute.

Method	PSNR	SSIM	FSIM	SRE	UIQ	BRISQUE	Parameters
Bicubic	29.1158	0.5186	0.4992	47.9362	0.3318	89.5672	-
SiGAN [10]	28.8599	0.5915	0.5298	47.7583	0.4317	42.5275	13.38M
AACNN [48]	29.8099	0.6105	0.5504	49.5380	0.4798	31.5483	2.49M
DDA-SRGAN [32]	29.3533	0.5507	0.5098	48.2359	0.3924	34.2122	1.10M
MASRGAN [33]	30.2477	0.6835	0.5853	50.2075	0.5522	43.9010	3.95M
SparNet [24]	29.1402	0.6201	0.5581	48.7875	0.4874	34.0678	25.44M
ATMFN [23]	29.4444	0.5964	0.5321	47.5626	0.4467	81.1625	64.26M
AGA-GAN (Ours)	30.8353	0.7323	0.6180	50.8118	0.6216	37.7595	9.77M
AGA-GAN+U-Net (Ours)	**31.2548**	**0.7655**	**0.6313**	**51.1707**	**0.6608**	**31.5656**	13.63M

5.3. Effect on downstream facial verification task

We further evaluate the performance of our proposed AGA-GAN+UNet against other baselines on downstream tasks like face verification. For this experiment we choose FaceNet-512 [46] to convert the images into a compact euclidean space. Hereafter, we compute the cosine similarity between the generated SR and the ground truth HR image using Equation 14.

$$\textit{cosine similarity}(\mathbf{t}, \mathbf{e}) = \frac{\mathbf{te}}{\|\mathbf{t}\|\|\mathbf{e}\|} = \frac{\sum_{i=1}^{n} \mathbf{t}_i \mathbf{e}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{t}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{e}_i)^2}} \tag{14}$$

Here, t and e represents the 512 dimensional embedding extracted by feeding HR and SR images into FaceNet-512. Now, cosine distance is calculated as $\textit{cosine distance} = 1 - \textit{cosine similarity}$, and the quantitative results are reported in Table 9. Since the SR and HR image have the same identity, in an ideal case there should no cosine distance between such a pair. The effectiveness of our proposed method in recovering salient facial structures allows AGA-GAN+UNet to generate SR images which are the most similar to their HR counterpart as shown in Table 9. AGA-GAN reports the highest similarity on the 4x upscaling task while AGA-GAN+UNet achieves the best performance on 8x upsampling on global and attribute specific images.

5.4. Effect of Feature Refinement using U-Net

We study the impact of adding a spatial and channel attention U-Net on the existing AGA-GAN. Using the prediction of AGA-GAN as input for U-Net, the network works on improving the facial features and generating finer and richer details to improve the overall visual quality. Performance gain

Table 9: Result comparison using average cosine distance between generated SR image and ground truth HR image.

Method	Upscale Factor		Attribute Categories					
	8x	**4x**	**Big nose**	**Eyeglasses**	**Goatee**	**Mouth**$^{\diamond}$	**Mustache**	**Eyes**$^{\diamond}$
Bicubic	0.8193	0.2461	0.7918	0.7982	0.7780	0.8361	0.7654	0.7980
SiGAN [10]	0.6271	0.8296	0.5916	0.6324	0.5883	0.6338	0.5640	0.6188
AACNN [48]	0.4522	0.1382	0.4531	0.4828	0.4327	0.4616	0.4335	0.4485
DDA-SRGAN [33]	0.4215	0.0936	0.4531	0.4828	0.4326	0.4616	0.4336	0.4485
MA-SRGAN [32]	0.4787	0.1452	0.4600	0.5062	0.4525	0.4811	0.4397	0.4652
SparNet [24]	0.5106	0.2565	0.4860	0.4938	0.4686	0.5062	0.4765	0.4882
ATMFN [23]	0.6545	0.3459	0.6172	0.6726	0.6311	0.6816	0.6112	0.6489
AGA-GAN	0.3727	**0.0833**	0.3311	0.3827	0.3561	0.3793	0.3562	0.3654
AGA-GAN+U-Net	**0.2866**	0.0905	**0.2775**	**0.2853**	**0.2601**	**0.2759**	**0.2595**	**0.2695**

* *The 'Mouth' and 'Eyes' refer to the 'Open Mouth' and 'Narrow eyes' attribute description*

observed from Table 1 and Table 2 show that U-Net is successful in refining and improving the quality of existing SR predictions by AGA-GAN. In Figure 9 we can observe the various artifacts generated in the SR predictions of AGA-GAN and their subsequent rectification and refinement by U-Net. This proves the scope of using U-Net for image refinement of candidate predictions by other face hallucination methods.

5.5. Evaluation when Attribute information is scarce

In real-world scenarios, it is not feasible that entire attribute information will be attainable with each LR image. To demonstrate the performance of AGA-GAN in the case when only partial attribute information is present we represent some known attributes is unknown. In Table 10, we observe that when 50% of the attributes are known, a drop in performance is noted which further drops when 75% of the attributes are unknown. The performance is still superior to other methods and demonstrates that AGA-GAN can be used in situations when only partial attribute information is present.

Table 10: Result comparison with partially unknown attribute sets.

Unknown attributes	**PSNR**	**SSIM**	**FSIM**	**SRE**	**UIQ**
50%	31.4353	0.7839	0.6496	52.4684	0.6670
75%	31.4266	0.7801	0.6450	52.4356	0.6661

Figure 9: Qualitative examples of feature refinement using U-Net.

6. Conclusion

In this work, we have presented our attribute-guided attention generative adversarial network (AGA-GAN) which utilizes attributes before an LR image to generate attention maps capable of progressively generating richer and spatially accurate high-resolution feature maps, eventually generating high-resolution predictions with high fidelity and rich features. We also explore the use of spatial and channel attention U-Net for refining and generating additional facial features in an existing SR prediction. Extensive experiments with five metrics display the superior performance of both AGA-GAN and AGA-GAN+U-Net for facial hallucination. Additionally, experiments with partially known attributes demonstrate the practicality of our method in real-world scenarios. Our future work will comprise of extending our AGA-GAN+U-Net framework for face anti-spoofing and face parsing.

7. Acknowledgement

This is a collaborative research work between Indian Statistical Institute, Kolkata, India and Østfold University College, Halden, Norway. The experiments in this paper were performed on a high performance computing platform "Experimental Infrastructure for Exploration of Exascale Computing" (eX3), which is funded by the Research Council of Norway.

References

[1] W. W. Zou, P. C. Yuen, Very low resolution face recognition problem, IEEE Transactions on image processing 21 (2011) 327–340.

[2] X. Ma, J. Zhang, C. Qi, Hallucinating face by position-patch, Pattern Recognition 43 (2010) 2224–2236.

[3] X. Zhang, X. Wu, Image interpolation by adaptive 2-d autoregressive modeling and soft-decision estimation, IEEE transactions on image processing 17 (2008) 887–896.

[4] J. Sun, Z. Xu, H.-Y. Shum, Gradient profile prior and its applications in image super-resolution and enhancement, IEEE Transactions on Image Processing 20 (2010) 1529–1542.

[5] K. Zhang, D. Tao, X. Gao, X. Li, Z. Xiong, Learning multiple linear mappings for efficient single image super-resolution, IEEE Transactions on Image Processing 24 (2015) 846–861.

[6] H. Chang, D.-Y. Yeung, Y. Xiong, Super-resolution through neighbor embedding, in: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004., volume 1, IEEE, 2004, pp. I–I.

[7] C. Dong, C. C. Loy, X. Tang, Accelerating the super-resolution convolutional neural network, in: European conference on computer vision, Springer, 2016, pp. 391–407.

[8] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, Z. Wang, Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1874–1883.

[9] X. Yu, F. Porikli, Ultra-resolving face images by discriminative generative networks, in: European conference on computer vision, Springer, 2016, pp. 318–333.

[10] C.-C. Hsu, C.-W. Lin, W.-T. Su, G. Cheung, Sigan: Siamese generative adversarial network for identity-preserving face hallucination, IEEE Transactions on Image Processing 28 (2019) 6225–6236.

[11] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.

[12] S. Baker, T. Kanade, Hallucinating faces, in: Proceedings Fourth IEEE international conference on automatic face and gesture recognition (Cat. No. PR00580), IEEE, 2000, pp. 83–88.

[13] C. Liu, H.-Y. Shum, C.-S. Zhang, A two-step approach to hallucinating faces: global parametric model and local nonparametric model, in: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, volume 1, IEEE, 2001, pp. I–I.

[14] B. K. Gunturk, A. U. Batur, Y. Altunbasak, M. H. Hayes, R. M. Mersereau, Eigenface-domain super-resolution for face recognition, IEEE transactions on image processing 12 (2003) 597–606.

[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] Y. Liang, X. Xie, J.-H. Lai, Face hallucination based on morphological component analysis, Signal Processing 93 (2013) 445–458.

[17] E. Zhou, H. Fan, Z. Cao, Y. Jiang, Q. Yin, Learning face hallucination in the wild, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 29, 2015.

[18] X. Chen, X. Wang, Y. Lu, W. Li, Z. Wang, Z. Huang, Rbpnet: An asymptotic residual back-projection network for super-resolution of very low-resolution face image, Neurocomputing 376 (2020) 119–127.

[19] H. Huang, R. He, Z. Sun, T. Tan, Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1689–1697.

[20] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, arXiv preprint arXiv:1406.2661 (2014).

[21] S. D. Indradi, A. Arifianto, K. N. Ramadhani, Face image super-resolution using inception residual network and gan framework, in: 2019 7th International Conference on Information and Communication Technology (ICoICT), IEEE, 2019, pp. 1–6.

[22] L. Yang, S. Wang, S. Ma, W. Gao, C. Liu, P. Wang, P. Ren, Hifacegan: Face renovation via collaborative suppression and replenishment, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1551–1560.

[23] K. Jiang, Z. Wang, P. Yi, G. Wang, K. Gu, J. Jiang, Atmfn: Adaptive-threshold-based multi-model fusion network for compressed face hallucination, IEEE Transactions on Multimedia 22 (2019) 2734–2747.

[24] C. Chen, D. Gong, H. Wang, Z. Li, K.-Y. K. Wong, Learning spatial attention for face super-resolution, IEEE Transactions on Image Processing 30 (2020) 1219–1231.

[25] K. Grm, W. J. Scheirer, V. Štruc, Face hallucination using cascaded super-resolution and identity priors, IEEE Transactions on Image Processing 29 (2019) 2150–2165.

[26] V. Mnih, N. Heess, A. Graves, K. Kavukcuoglu, Recurrent models of visual attention, arXiv preprint arXiv:1406.6247 (2014).

[27] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proc. of Comput. Vis. and Patt. Recogn., 2018, pp. 7132–7141.

[28] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: International conference on machine learning, PMLR, 2015, pp. 2048–2057.

[29] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang, Residual attention network for image classification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3156–3164.

[30] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.

[31] J. Fu, H. Zheng, T. Mei, Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4438–4446.

[32] C.-E. Huang, Y.-H. Li, M. S. Aslam, C.-C. Chang, Super-resolution generative adversarial network based on the dual dimension attention mechanism for biometric image super-resolution, Sensors 21 (2021) 7817.

[33] C.-E. Huang, C.-C. Chang, Y.-H. Li, Mask attention-srgan for mobile sensing networks, Sensors 21 (2021) 5973.

[34] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, Y. Fu, Residual dense network for image super-resolution, in: Proc. of Comput. Vis. and Patt. Recogn., 2018.

[35] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Proc. of AAAI Conf. Artifi. Intelli., volume 31, 2017.

[36] B. Lim, S. Son, H. Kim, S. Nah, K. Mu Lee, Enhanced deep residual networks for single image super-resolution, in: Proc. of Comput. Vis. and Patt. Recogn. Worksh., 2017, pp. 136–144.

[37] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.

[39] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training gans, arXiv preprint arXiv:1606.03498 (2016).

[40] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.

[41] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Processing Letters 23 (2016) 1499–1503.

[42] L. Zhang, L. Zhang, X. Mou, D. Zhang, Fsim: A feature similarity index for image quality assessment, IEEE transactions on Image Processing 20 (2011) 2378–2386.

[43] C. Lanaras, J. Bioucas-Dias, S. Galliani, E. Baltsavias, K. Schindler, Super-resolution of sentinel-2 images: Learning a globally applicable deep neural network, ISPRS Journal of Photogrammetry and Remote Sensing 146 (2018) 305–319.

[44] Z. Wang, A. C. Bovik, A universal image quality index, IEEE signal processing letters 9 (2002) 81–84.

[45] A. Mittal, A. K. Moorthy, A. C. Bovik, No-reference image quality assessment in the spatial domain, IEEE Transactions on image processing 21 (2012) 4695–4708.

[46] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815–823.

[47] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: Proceedings of International Conference on Computer Vision (ICCV), 2015.

[48] C.-H. Lee, K. Zhang, H.-C. Lee, C.-W. Cheng, W. Hsu, Attribute augmented convolutional neural network for face hallucination, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2018, pp. 721–729.

[49] N. C. Rakotonirina, A. Rasoanaivo, Esrgan+: Further improving enhanced super-resolution generative adversarial network, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 3637–3641.