Journal of the Brazilian Computer Society, 2026, 32:1, doi: 10.5753/jbcs.2026.5884

Robust Face Super-Resolution and Recognition Through Multi-Feature Aggregation in Diffusion Models

Marcelo dos Santos [**Federal University of Paraná** | ***msantos@inf.ufpr.br***]
Rayson Laroca [**Pontifical Catholic University of Paraná, Federal University of Paraná**| ***rayson@ppgia.pucpr.br***]
João Carlos Raposo Neves [**University of Beira Interior** | ***jcneves@ubi.pt***]
David Menotti [**Federal University of Paraná** | ***menotti@inf.ufpr.br***]

Federal University of Paraná, Department of Informatics, Curitiba-PR, 81531-970, Brazil

Received: 11 April 2025 • **Accepted:** 23 April 2026 • **Published:** 29 May 2026

Abstract. Images acquired in surveillance environments often suffer from conditions such as low resolution, variations in pose, irregular illumination, and occlusions. Due to the low quality of these images, face recognition algorithms often struggle. This major limitation can be addressed by employing super-resolution techniques that enhance the details of the image. However, due to the high degree of difficulty of the problem, most super-resolution algorithms tend to cause distortions in the image and in the individual's identity. Thus, additional information must be incorporated into the processing to improve recognition robustness. In this regard, surveillance cameras can capture multiple images, even at low quality, and the data extracted from these images, such as consecutive video frames, can significantly enhance both super-resolution and facial recognition. In this work, we introduce FASR++, a diffusion-model-based super-resolution algorithm. It leverages a reference low-resolution image and features extracted from multiple auxiliary low-quality images to generate a super-resolved output, minimizing distortions in the individual's identity. Our approach recovers facial features without explicitly providing soft attributes or computing a function gradient to guide the reconstruction process. FASR++ generates high-quality images that can considerably improve performance in face recognition tasks when used as a pre-processing step. We validate our approach on two standard face recognition datasets and attain state-of-the-art results for verification, face recognition, and image quality metrics such as PSNR, SSIM, and LPIPS.

Keywords: Diffusion models, Super-Resolution, Face Recognition

1 Introduction

In surveillance scenarios, the presence of noise, occlusions, variations in illumination, and varying poses poses a challenge even for state-of-the-art (SOTA) face recognition algorithms, leading to a significant decline in performance [Zhu *et al*., 2016]. The most critical issue is the low-resolution and low-quality images acquired in real-world scenarios. Thus, the idea of utilizing super-resolution (SR) algorithms as a pre-processing step for face recognition is not new [Bilgazyev *et al*., 2011] and arose as a natural solution to generate images with higher quality. However, the super-resolution problem is inherently ill-posed, making the recovery of fine details and a reliable identity quite challenging [Baker and Kanade, 2002; Jiang *et al*., 2021; Nascimento *et al*., 2022, 2024]. In this sense, some works have focused on performing super-resolution by leveraging soft attributes such as eyeglasses, beards, mustaches, gender, and others as an additional source of information to reduce ambiguity and provide more robust results [Lee *et al*., 2018; Yu *et al*., 2018; Lu *et al*., 2018; dos Santos *et al*., 2024b]. Nevertheless, facial attributes are often indistinct in low-resolution (LR) images, making reliable identification challenging. Also, obtaining these attributes requires a classifier or manual extraction, which is not very efficient [dos Santos *et al*., 2024b]. However, many characteristics can be employed to assist super-resolution algorithms. These include subtle facial proportions, skin textures, shapes, and other high-level, more abstract features that are not easily labeled or categorized.

Given the scarcity of robust super-resolution methods capable of preserving identity, this work aims to address this gap by focusing on identity preservation. We develop *Feature Aggregation Super-Resolution* (FASR++), a robust SR algorithm that recovers crucial features for face recognition. It is more effective because, in addition to the LR image, it also takes as input a reliable vector of facial features derived from a set of LR images, such as a series of video frames or independent photos of an individual. This new vector has a higher signal-to-noise ratio than each individual vector. We incorporate it into the network, merging its information with the LR image to generate an SR version. In this way, our algorithm effectively recovers facial information from an image, yielding higher-quality results with minimal distortion of identity. Notably, FASR++ employs a diffusion model based on a Stochastic Differential Equation (SDE) and does not require a classifier to guide the reverse diffusion process.

A preliminary version of this work, introducing the Feature Aggregation Super-Resolution (FASR) method, was published at the 2024 Conference on Graphics, Patterns and Images (SIBGRAPI) [dos Santos *et al*., 2024a]. This paper builds upon that work with the following key improvements: (i) we develop a neural network specifically designed to enhance feature integration in low-resolution images, enabling more effective fusion and recovery of high-frequency details. A hypoth-

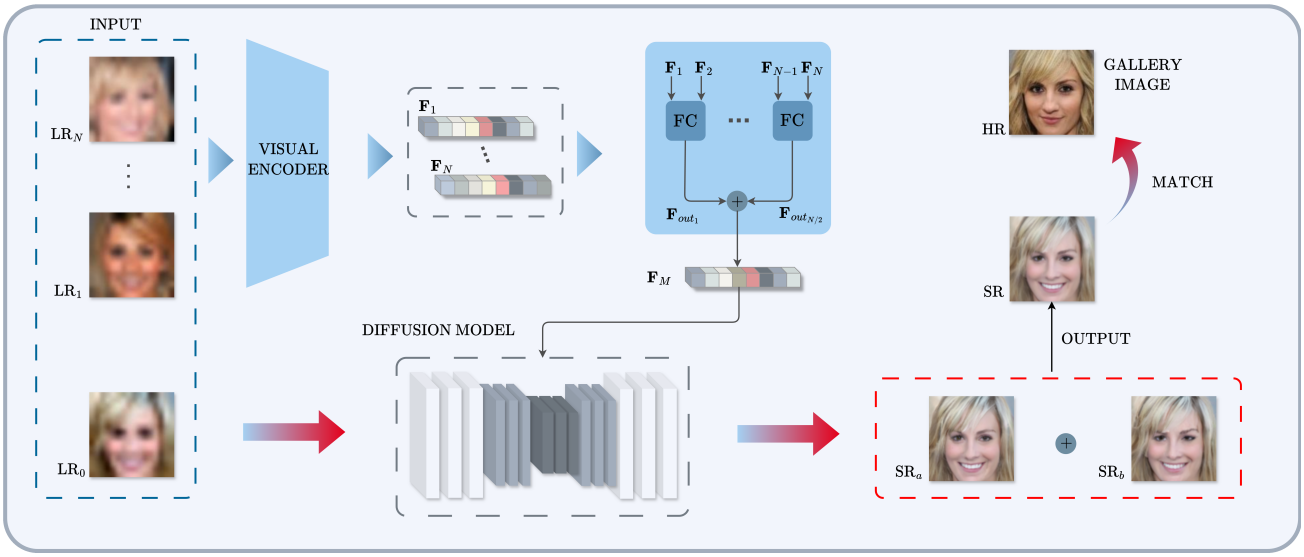

Figure 1. Overview of the proposed method. At inference time, a set of $N+1$ low-resolution images is collected from an individual. The images $\text{LR}_1, \ldots, \text{LR}_N$ are processed to extract feature representations $\mathbf{F}_1, \ldots, \mathbf{F}_N$, which are then aggregated through an ensemble of Feature Combiner (FC) modules to produce the merged feature vector $\mathbf{F}_M$. The reference image LR_0 is jointly integrated with $\mathbf{F}_M$ into the diffusion model to generate two super-resolved outputs, SR_a and SR_b. Their arithmetic mean yields the final super-resolved image SR, which is compared against a gallery of facial images for identity matching.

esis test validates its ability to merge complementary features; (ii) we incorporate additional evaluation metrics, expanding beyond recognition and verification performance to include image quality assessments such as Structural Similarity Index Measure (SSIM), Peak Signal-to-Noise Ratio (PSNR), and Learned Perceptual Image Patch Similarity (LPIPS) [Zhang *et al*., 2018]; (iii) we conduct a comprehensive ablation study, demonstrating that the proposed method significantly outperforms the original FASR and other baseline models.

Our method's effectiveness has been validated on the popular CelebA [Liu *et al*., 2015] and Quis-Campi [Neves *et al*., 2018] datasets, and the key contributions are as follows:

- We present a neural network that fuses and enhances low-resolution features, and we demonstrate its effectiveness in feature integration;
- We introduce FASR++, an improvement of FASR [dos Santos *et al*., 2024a], which utilizes the proposed neural network to generate high-quality merged features for assisting the reconstruction of super-resolution images in the diffusion model;
- Our approach achieves superior qualitative results, producing more natural images with less distortion compared to SOTA super-resolution algorithms;
- Our quantitative results outperform SOTA algorithms in both image quality metrics, such as PSNR, SSIM and LPIPS, and identity-related metrics, including Area Under the Curve (AUC) in 1:1 verification, and accuracy in the 1:N identification.

2 Related Work

Sohl-Dickstein *et al*. [2015] introduced a generative model based on principles from non-equilibrium thermodynamics in their seminal work. Two other influential studies in the field of diffusion models are Denoising Diffusion Probabilistic Models (DDPMs) [Ho *et al*., 2020] and Score-Based Generative Models (SGMs) [Song and Ermon, 2019, 2020]. In [Song *et al*., 2021], DDPM and SGM are generalized for continuous time steps and noise levels using Stochastic Differential Equations (SDEs), expanding the range of research possibilities in diffusion models.

Due to the rapid evolution of diffusion models, various opportunities for their application have emerged. Recent works include the generation of audio, graphs, and shapes, as well as image synthesis, solutions of general inverse problems, and applications in medical images [Niu *et al*., 2020; Cai *et al*., 2020; Ho *et al*., 2020; Song and Ermon, 2019; Song *et al*., 2021, 2022]. The full potential of diffusion models can also be leveraged through multi-domain data integration, such as text-to-image translation [Saharia *et al*., 2022] and image editing [Zhang *et al*., 2023]. Additionally, Richter *et al*. [2023] combines audio-visual information for speech enhancement.

SR is another important application of diffusion models and is investigated in this work. In Saharia *et al*. [2023], an adaptation of the DDPM model produces high-quality SR images. Similarly, SRDiff [Li *et al*., 2022] employs diffusion models to estimate the difference between the original LR image and a high-resolution (HR) image, resulting in an SR image. In [dos Santos *et al*., 2022], SDEs were used to generate SR images. Additionally, dos Santos *et al*. [2024b] performs SR by incorporating attribute information such as beard, gender, and the presence of eyeglasses to generate high-quality images. However, their approach has the drawback that these attributes must be explicitly provided to the algorithm, which cannot be easily estimated in LR images.

In [Suin *et al*., 2024], an identity-preserving SR method was developed. In both [dos Santos *et al*., 2024b] and [Suin *et al*., 2024], a gradient must be calculated during the image reconstruction phase, which can increase computational cost. In this study, we develop an algorithm that restores image attributes by supplying a compact descriptor of facial features

for the algorithm.

Despite the impressive results achieved by diffusion models, their primary drawback is the high execution time caused by their iterative nature. Nevertheless, this issue is expected to be mitigated in the near future, as many studies focus on improving the computational efficiency of these methods. For a more in-depth discussion on accelerating sampling and enhancing efficiency in diffusion models, refer to [Jolicoeur-Martineau *et al*., 2021; Vahdat *et al*., 2021; Meng *et al*., 2023].

3 Proposed Method

In this section, we present the general concept of the proposed method, followed by the description of a Feature Combiner module, the theoretical background on diffusion models formulated as SDEs, the model architecture, and the conditioning mechanisms based on low-resolution images, time, and feature embeddings.

3.1 General Idea

As previously noted, images captured in surveillance environments are often of low quality. Nevertheless, in certain instances, a video of a particular person can provide multiple low-resolution images that, when combined, can increase the valuable information necessary to recognize an individual.

In this work, we employ an SR algorithm to enhance a low-resolution image (LR_0), recovering useful information for face recognition (see Figure 1). We use a set of low-resolution auxiliary images $\mathrm{LR}_1, \ldots, \mathrm{LR}_N$ of the same individual to extract a set of compact descriptors $\mathbf{F}_1, \ldots, \mathbf{F}_N$. We train a Feature Combiner (FC) module (see Figure 2) to merge two vectors and recover the image's high-frequency information. The vectors $\mathbf{F}_1, \ldots, \mathbf{F}_N$ are combined through an ensemble of FCs to generate a representative compact descriptor $\mathbf{F}_M$. The reference low-resolution image LR_0 and the merged vector $\mathbf{F}_M$ are input into a diffusion model, which generates super-resolution images with minimal distortions in the identity. Due to the stochasticity of diffusion models and the ill-posed nature of super-resolution problems, we can generate different solutions with small variations consistent with the same input LR_0. To increase the method's robustness, we generate two super-resolution images, SR_a and SR_b, and combine them through an average to generate the final image, SR. This general framework leads to enhanced reliability concerning person identification. Moreover, the algorithm successfully retrieves high-level features that might not be clearly visible but significantly enhance recognition accuracy and image quality.

3.2 Feature Combiner (FC)

To train the Feature Combiner (FC) module, we need a collection of low-resolution images and their high-resolution versions. Specifically, for a given identity, let $\mathrm{LR}_1, \ldots, \mathrm{LR}_N$ denote the set of N low-resolution images, and let $\mathrm{HR}_1, \ldots, \mathrm{HR}_N$ denote the respective set of high-resolution images. We then extract a set of compact face descriptors from these images using a pre-trained visual encoder to obtain the descriptors $\mathbf{F}_1, \ldots, \mathbf{F}_N, \mathbf{F}_1^{\mathrm{HR}}, \ldots, \mathbf{F}_N^{\mathrm{HR}}$ respectively. The feature vectors from low-resolution images are combined in pairs using an ensemble of FC modules and posteriorly averaged to obtain an approximation of a reliable descriptor of a high-resolution image.

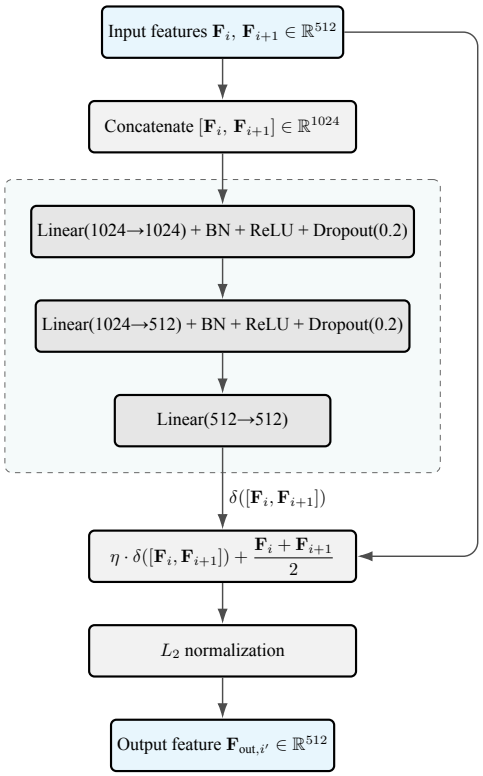

Figure 2. Architecture of the Feature Combiner. The input features $\mathbf{F}_i$ and $\mathbf{F}_{i+1}$ are concatenated and processed by the neural network δ, which is composed of two fully connected layers with Batch Normalization (BN), ReLU activation, and Dropout (0.2), followed by a final linear layer. The resulting output of δ is combined with the mean of the input features to generate the final merged representation $\mathbf{F}_{\mathrm{out},i'}$.

More specifically, given two feature vectors $\mathbf{F}_i$ and $\mathbf{F}_{i+1}$, the FC module takes them as input and produces a single merged representation $\mathbf{F}_{\mathrm{out},i'}$, where $i' = (i+1)/2$, as illustrated in Figure 2. The FC module consists of the mean of the input features and a refinement network δ. The mean operation was adopted as a baseline because the input low-resolution images, and consequently their extracted features, often contain noise and small distortions. Averaging allows the model to amplify the facial components that are common across different images of the same identity while attenuating uncorrelated noise and spurious variations. The FC then combines this mean with a learnable correction term generated by the neural network δ, which captures nonlinear relationships between the two feature vectors, modeling complex dependencies that cannot be represented by simple averaging. Formally, the operation is

$$\mathbf{F}_{\mathrm{out},i'} = \mathrm{FC}\left(\mathbf{F}_i, \mathbf{F}_{i+1}\right) = \frac{\mathbf{F}_i + \mathbf{F}_{i+1}}{2} + \eta \cdot \delta\left(\mathbf{F}_i, \mathbf{F}_{i+1}\right), \tag{1}$$

where $\delta(\cdot)$ denotes the neural network and $\eta \in \{0, 1\}$ is a control parameter used to assess the influence of δ on the final results.

The architecture of δ, presented in Figure 2, consists of three fully connected layers. The input is formed by concate-

nating the feature vectors $\mathbf{F}_i$ and $\mathbf{F}_{i+1}$, producing a 1024-dimensional representation. This concatenated vector is processed by two sequential linear transformations, each followed by Batch Normalization (BN) and ReLU activation, with a dropout rate of 0.2 applied after each nonlinear operation to improve generalization and reduce overfitting. The chosen dropout rate of 0.2 was determined through validation experiments and aligns with the typical values reported by Labach *et al*. [2019]. A final linear transformation produces a 512-dimensional feature refinement, which is combined with the mean $(\mathbf{F}_i + \mathbf{F}_{i+1})/2$ to generate the fused representation $\mathbf{F}_{\text{out},i'}$. Finally, an L_2 normalization is applied to ensure unit-length embeddings.

To optimize the network parameters, we employ a *triplet loss*, aiming to maximize the similarity between positive matches while minimizing it for negative ones. In this framework, the anchor, positive and negative samples are:

- **Anchor**: $\mathbf{F}_{\text{out},i'} = \text{FC}\,(\mathbf{F}_i, \mathbf{F}_{i+1})$, the combined feature representation generated by the FC module;
- **Positive sample**: The arithmetic mean of feature vectors from the respective HR image, i.e.:

$$\frac{\mathbf{F}_i^{\text{HR}} + \mathbf{F}_{i+1}^{\text{HR}}}{2}; \tag{2}$$

- **Negative sample**: A feature vector from a HR image of a different identity.

After the network δ is trained and each FC module effectively merges two feature vectors, an ensemble of FCs is employed to obtain the overall merged representation for a given individual. The final feature vector is computed by averaging all $N/2$ merged outputs $\mathbf{F}_{\text{out},i'}$, resulting in the final merged feature $\mathbf{F}_M$, as expressed by

$$\mathbf{F}_M = \frac{2}{N}\sum_{i'=1}^{N/2}\mathbf{F}_{\text{out},i'} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{F}_i + \frac{2\eta}{N}\sum_{i=1}^{N/2}\delta(\mathbf{F}_{2i-1}, \mathbf{F}_{2i}), \tag{3}$$

for even N. When N is odd, the last feature vector $\mathbf{F}_N$ is averaged together with the $(N-1)/2$ merged features $\mathbf{F}_{\text{out},i'}$. This formulation is general and applies to any number of input images. The averaging operation in Equation 3 can be interpreted as analogous to the average pooling mechanism commonly employed in convolutional neural networks, where feature-wise averaging effectively reduces noise while preserving the dominant structural information.

In Equation 3, other operations, such as the maximum or the sum, could also be used. However, based on experimental analysis, we found that the mean yields superior results compared to these alternatives. Using the maximum may distort the features by emphasizing individual noisy components rather than the common underlying structure. A simple sum, in turn, would unnecessarily amplify the magnitude of the features as N increases. In contrast, the mean provides a stable aggregation that yields merged features $\mathbf{F}_{\text{out},i'}$ closer to the target high-resolution ground truths, ensuring a coherent estimate of the underlying representation. Since the individual feature vectors are normalized, the mean effectively balances their contributions without being dominated by any single component. Moreover, residual noise components may still be present, and averaging serves as a natural denoising mechanism that increases the signal-to-noise ratio and stabilizes the overall representation across different image conditions.

From Equation 3, it follows that the mean of $\mathbf{F}_{\text{out},i'}$ is equivalent to a single global mean over all input features $\mathbf{F}_i$, plus an additional term involving $\delta(\cdot,\cdot)$ that represents the nonlinear refinement computed from sequential feature pairs by the Feature Combiner. Therefore, if another nonlinear function were more suitable for feature aggregation than the mean, the δ network would implicitly compensate for it.

This formulation ensures that the learned representation effectively consolidates the features while preserving identity and enhancing discriminability among individuals.

3.3 Theoretical Background

In the context of image generation, diffusion models have two phases: forward diffusion and reverse diffusion. During forward diffusion, Gaussian noise is added to the image, and a network is trained to predict this noise. In reverse diffusion, an image composed purely of noise is iteratively denoised and transformed into an image that follows a distribution similar to the images in the training set. If the diffusion procedure is continuous, it can be modeled using an SDE.

According to Song *et al*. [2021]; Anderson [1982], a forward diffusion process $\{\mathbf{x}(t)\}_{t=0}^{T}$ and its reverse are, respectively, modeled using the following SDEs:

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}, \tag{4}$$

$$\mathrm{d}\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g(t)^2\nabla_{\mathbf{x}}\log p_t(\mathbf{x})]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}, \tag{5}$$

where $\mathbf{f}(\mathbf{x}, t)$ is the drift coefficient, $g(t)$ is a diffusion coefficient, $\mathbf{w}$ and $\bar{\mathbf{w}}$ are Wiener process (the latter runs backward in time) and p_t is the probability density of $\mathbf{x}(t)$. Kloeden and Platen [2011]; Särkkä and Solin [2019] supply more details about Itô SDEs and the Wiener process.

Here, we consider $\mathbf{x}_t$ as an image to be denoised. At $t = 0$, the noise level in the image is zero, and at $t = T$, the noise is at its maximum, and there is no information on the image. To obtain a super-resolved image, we solve the reverse diffusion process defined in Equation 5. For this purpose, a deep neural network s_θ is employed to approximate the score function $\nabla_{\mathbf{x}}\log p_t(\mathbf{x})$. During this reverse process, s_θ is conditioned on both the reference low-resolution image LR_0, denoted by $\mathbf{y}$, and the merged feature vector $\mathbf{F}_M$, which provides complementary guidance.

The training of the neural network s_θ is achieved by optimizing the following loss function [Vincent, 2011]:

$$\min_\theta \mathbb{E}_{t\sim\mathcal{U}[0,T]}\mathbb{E}_{\mathbf{x}(0)\sim p(\mathbf{x}(0))}\mathbb{E}_{\mathbf{x}(t)\sim p_t(\mathbf{x}(t)|\mathbf{x}(0))}\big[\lambda(t) \times \|s_\theta(\mathbf{x}(t), \mathbf{y}, \mathbf{F}_{\mathbf{y}}^{\text{HR}}, t) - \nabla_{\mathbf{x}(t)}\log p(\mathbf{x}(t)|\mathbf{x}(0))\|_2^2\big], \tag{6}$$

where $\lambda(t)$ is a positive weighting function, $p(\mathbf{x}(t)|\mathbf{x}(0))$ is the transition kernel from $\mathbf{x}(0)$ to $\mathbf{x}(t)$ and $\mathbf{F}_{\mathbf{y}}^{\text{HR}}$ represents the ground-truth features extracted from the high-resolution version of $\mathbf{y}$.

Here, we use the Variance Exploding (VE) case described in [Song *et al*., 2021] with $\mathbf{f}(\mathbf{x}, t)$ and $g(t)$ given respectively by:

$$\mathbf{f}(\mathbf{x}, t) = \mathbf{0}, \quad g(t) = \sqrt{\frac{\mathrm{d}\sigma^2(t)}{\mathrm{d}t}}, \tag{7}$$

where $\sigma(t) = \sigma_{min}\,(\sigma_{max}/\sigma_{min})^t$ denotes the noise level of the image at the time t.

For $\mathbf{f}(\mathbf{x}, t)$ and $g(t)$ described above, the mean and variance of $p(\mathbf{x}(t)|\mathbf{x}(0))$ are given by [Song *et al*., 2021]:

$$\boldsymbol{\mu}(t) = \mathbf{x}(0),\ \boldsymbol{\Sigma}(t) = [\sigma^2(t) - \sigma^2(0)]\mathbf{I}. \tag{8}$$

Thus, we can analytically compute $\nabla_{\mathbf{x}} \log p(\mathbf{x}(t)|\mathbf{x}(0))$ in Equation 6, allowing for efficient model training. Once the network well estimates the gradient, we generate an SR image $\mathbf{x}(0)$ by changing $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ by $s_\theta(\mathbf{x}(t), \mathbf{y}, \mathbf{F}_M, t)$ in the reverse process (Equation 5) and solving it from $t = T$ to $t = 0$ using the Euler-Maruyama method [Kloeden and Platen, 2011; Särkkä and Solin, 2019].

3.4 Diffusion Model Architecture

The proposed model adopts the NCSN++ [Song *et al*., 2021] backbone, implemented as a U-Net [Ronneberger *et al*., 2015] with residual and attention blocks distributed in multiple resolutions. The network comprises seven resolution levels ($128 \to 64 \to 32 \to 16 \to 8 \to 4 \to 2$), with channel widths scaled by $(1, 1, 2, 2, 2, 2, 2)$ relative to an initial number of 128 feature maps. Each level contains two residual blocks, followed by either downsampling or upsampling operations between successive resolutions.

Each residual block follows a standard design consisting of Group Normalization, a SiLU nonlinearity, and two 3×3 convolutional layers. A dropout rate of 0.1 is applied within each block to improve regularization, while a 1×1 convolution is used for channel projection whenever the input and output dimensions differ. In addition, self-attention layers are inserted at the 16×16 resolution to capture long-range spatial dependencies across feature maps [Vaswani *et al*., 2017].

The network is conditioned on three signals: the low-resolution input image, the time embedding, and the merged feature vector. In each residual block, temporal and feature signals are injected through independent dense layers after SiLU activation, as described in Subsection 3.5.

Finally, a 3×3 convolution maps the feature maps back to the RGB space. This configuration, combining residual, attention and multi-scale conditioning mechanisms, improves representational capacity and training stability [He *et al*., 2016; Wang *et al*., 2017; Liang *et al*., 2021].

3.5 Model Conditioning

To guide the diffusion process, our model employs three complementary conditioning mechanisms: (i) the low-resolution image, which provides spatial and textural priors; (ii) the time embedding, which encodes the current denoising stage; and (iii) the merged feature vector, which conveys high-level identity information obtained from the ensemble of feature combiners. Each of these conditioning signals plays a distinct role in steering the generation process toward faithful and identity-preserving reconstructions.

LR Image Conditioning. To condition the model on low-resolution input, we follow a strategy similar to that adopted in [dos Santos *et al*., 2022; Saharia *et al*., 2023], where the conditioning image is directly concatenated with the noisy input image along the channel dimension. Specifically, the low-resolution image $\mathbf{y}$ and the noisy image $\mathbf{x}_T$ (the sample being progressively denoised) are concatenated to form a six-channel tensor

$$\text{concatenate}[\mathbf{y}, \mathbf{x}_T] \in \mathbb{R}^{6\times H\times W},$$

where H and W denote the spatial dimensions of the image. This tensor serves as the input to the U-Net backbone, allowing the network to take advantage of spatial and textural cues from the LR image throughout the denoising process.

Time Conditioning. The temporal conditioning follows the strategy adopted in other diffusion models [Song *et al*., 2021], where the diffusion timestep $t \in [0, 1]$ is mapped into a high-dimensional embedding using Gaussian Fourier features [Tancik *et al*., 2020]. For a given timestep t, the Fourier mapping is computed as

$$\mathbf{E}_t = \text{concatenate}[\sin(2\pi\boldsymbol{\omega}t),\ \cos(2\pi\boldsymbol{\omega}t)], \tag{9}$$

where $\boldsymbol{\omega} \in \mathbb{R}^d$ is a fixed, non-trainable vector of random frequencies drawn from a normal distribution. This produces a time embedding $\mathbf{E}_t \in \mathbb{R}^{2d}$ (with $d = 256$ in our implementation, yielding 512 dimensions).

The embedding $\mathbf{E}_t$ is then processed by a SiLU activation followed by a single linear layer that projects it into the same channel dimension C as the U-Net feature maps. To enable element-wise addition with the convolutional features, $\mathbf{E}_t$ is reshaped to $\mathbb{R}^{C\times 1\times 1}$ and broadcast across H and W, ensuring that the same temporal modulation is applied uniformly to all spatial positions within each residual block.

Time conditioning plays a crucial role in diffusion models, as it allows the network to adapt its denoising behavior according to the current noise level along the diffusion trajectory, improving both temporal coherence and reconstruction quality.

Feature Conditioning. Analogously, the model is also conditioned on the merged feature vector $\mathbf{F}_M$ obtained from the ensemble of feature combiners. This vector is processed through a SiLU activation followed by a single linear layer, projecting it into the same channel dimension C as the network feature maps. Then it is reshaped to $\mathbb{R}^{C\times 1\times 1}$ and broadcast across H and W to allow element-wise addition with the U-Net feature maps. In this way, the conditioning provided by $\mathbf{F}_M$ acts as a global control signal, influencing all spatial locations uniformly while preserving the channel-wise semantics learned by the network.

Integration within Residual Blocks. We now describe how these conditioning signals are incorporated into the network. At each residual block of the U-Net, let h denote the intermediate feature map after normalization and activation. The temporal embedding and the merged feature vector are injected into these residual blocks through independent linear layers D_t and D_f, as follows:

$$h \leftarrow h + \mathrm{D}_t(\mathrm{SiLU}(\mathbf{E}_t))_{\text{reshaped}} + \mathrm{D}_f(\mathrm{SiLU}(\mathbf{F}_M))_{\text{reshaped}}. \tag{10}$$

The layers D_t and D_f project these embeddings into the same channel dimension as h, and the resulting tensors are reshaped to match the spatial dimensions of the feature maps. This mechanism allows both the diffusion timestep and the identity-related information to condition the U-Net feature maps across multiple scales, as illustrated in Figure 3.

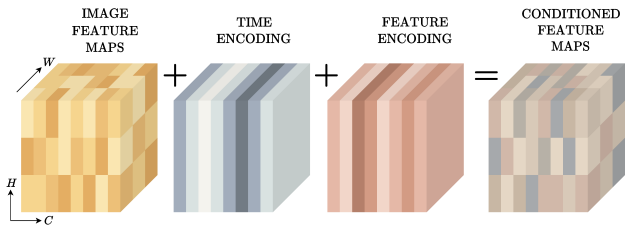

Figure 3. Time and features encoding. Illustration of the conditioning mechanism in the proposed diffusion model. The first block represents the image feature maps of shape (B, C, H, W), where B is the batch size. In this example, we illustrate a single element from the batch for clarity. The time encoding and feature encoding tensors are broadcast along the spatial dimensions so that their values remain constant across H and W, enabling element-wise addition with the image feature maps. The final block represents the feature maps conditioned on time and feature embeddings, which are subsequently processed by the following layers of the network.

Qualitative Demonstration of Feature Conditioning. To demonstrate the effectiveness of using the feature vector to generate an SR image, we trained our model with $\mathbf{y} = 0$ in Equation 6, i.e., without using the LR image, and employed the ground-truth feature vectors to produce images, as illustrated in Figure 4. These images demonstrate the algorithm's ability to reconstruct high-level features that encode identity-specific semantics, including facial geometry, landmark configuration, and characteristic texture patterns. The low-resolution image, in turn, provides the global spatial layout and low-frequency information required to preserve geometric coherence, pose, and illumination consistency during reconstruction. It defines the structural basis for the generative process. When both signals are combined, they enable the model to generate visually convincing and identity-consistent faces.

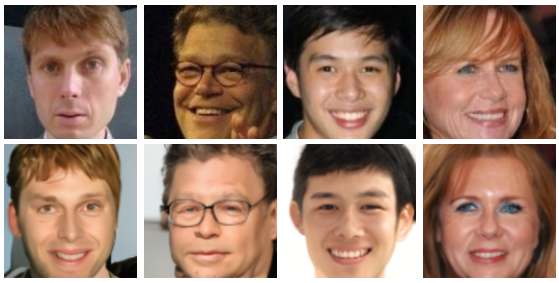

Figure 4. Efficacy of FASR++ for face reconstruction. Extracted from [dos Santos *et al*., 2024a]. First row: original HR images from the CelebA dataset [Liu *et al*., 2015]. Second row: synthetic HR images generated solely from the feature vectors extracted from the corresponding images in the first row.

4 Experiments and Results

This section describes the experimental setup and the results obtained on two different datasets. Lastly, we present the ablation study and examine some extreme cases where the algorithm may fail.

4.1 Experiments

In this study, we explored four datasets: FFHQ [Karras *et al*., 2019], CASIA-WebFace [Yi *et al*., 2014], CelebA [Liu *et al*., 2015], and Quis-Campi [Neves *et al*., 2018], the latter of which originates from surveillance scenarios. The FFHQ dataset was used for the diffusion model training, with 10^6 training steps performed, while the CASIA-WebFace dataset was specifically used to train the δ network. Further details on the δ training procedure are provided in the next section. CelebA was used to test our approach, with 500 identities selected. Each identity comprises multiple images, with one randomly chosen as the gallery image. A second image is downsampled to create an LR probe image. The remaining images were also downsampled and used to extract features, assisting the reconstruction of the LR probe image.

A complementary test to further validate our algorithm was conducted on a real-world scenario from the Quis-Campi dataset, where the images pose additional challenges for SR and face recognition algorithms [Neves *et al*., 2018]. We selected 90 identities and used five downsampled images as probe images for each identity. These images were then used to calculate an average feature vector, which was utilized to support the generation of the SR image. In addition, the dataset already contains gallery images obtained in a controlled environment for each identity.

The parameters controlling the noise level over time were set at $\sigma_{min} = 0.001$ and $\sigma_{max} = 348$. We worked with images of 128×128 pixels. For producing LR images, we applied 8×8 downsampling followed by upsampling using bicubic interpolation to achieve a final size of 128×128 pixels. We used 2,000 steps to solve the SDE for image reconstruction.

The feature vector used for both training the SR algorithm and facial recognition consists of a 512-dimensional vector generated through AdaFace [Kim *et al*., 2022] with a ResNet backbone [He *et al*., 2016] trained on the CASIA-WebFace dataset [Yi *et al*., 2014]. Image descriptors were compared using the cosine similarity metric. For the recognition task, we compare the SR-recovered images against the gallery images. Our proposed algorithm is compared against SOTA algorithms: SPARNET [Chen *et al*., 2020], GFPGAN [Wang *et al*., 2021], SwinIR [Liang *et al*., 2021], SDE-SR [dos Santos *et al*., 2022], IDM [Gao *et al*., 2023], SR3 [Saharia *et al*., 2023], and SRDG [dos Santos *et al*., 2024b].

4.2 Training and Evaluation of the δ Network

To train the δ network, we used the CASIA-WebFace dataset [Yi *et al*., 2014]. In our experiments a total of 490,623 images from 10,572 identities were used, corresponding to an average of 46.41 images per subject. The dataset was divided into training (85%) and validation (15%) subsets. The model was trained for up to 20 epochs, with early stopping applied when no improvement in the validation loss was observed for five consecutive epochs.

To compute the triplet loss, two high-resolution images were randomly selected for each identity. The mean of their feature vectors was used as the positive sample, while the features extracted from their corresponding low-resolution versions were used as inputs to the Feature Combiner module to generate the anchor representation. The negative sample was obtained from the high-resolution features of a different identity.

The work of Yuan *et al*. [2020] shows that training can collapse when the margin is too large relative to the initial embedding spread, preventing the embeddings from separating properly. Following this observation, the margin was empirically set slightly above the average gap between the anchor–positive and anchor–negative distances to enhance training stability, resulting in a final value of 0.495.

To demonstrate the effectiveness of the proposed ensemble of FC modules in merging low-resolution features, we conducted experiments on the CelebA dataset using the identities selected in Section 4.1. For each identity, the features of the low-resolution images $\mathrm{LR}_1, \ldots, \mathrm{LR}_N$ are fused using the ensemble of FCs, and the resulting feature representation is compared to the corresponding ground truth gallery feature via cosine similarity. We considered two approaches: (i) $\eta = 0$, where only the arithmetic mean is used for merging low-resolution features, and (ii) $\eta = 1$, where the network δ refines the arithmetic mean during feature integration. This comparison highlights the advantage of using the network δ for refinement over a simple averaging scheme.

The results show that the mean similarity score increased from 0.162 for $\eta = 0$ to 0.350 for $\eta = 1$, representing a 116% improvement. A paired t-test confirmed that this difference is highly significant ($t = -46.33$, $p < 10^{-12}$), demonstrating that the proposed δ network effectively enhances feature merging and produces representations that more closely match the high-resolution gallery features. The histogram analysis in Figure 5 further supports this conclusion, showing a clear rightward shift in the score distribution when the δ network is applied to refine the merging process. This improvement in the feature space ultimately translates into better face recognition performance, as discussed in Section 4.3.

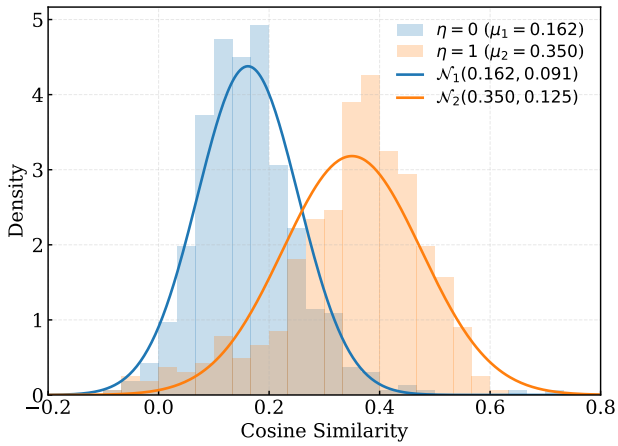

Figure 5. Histogram of the similarity score distributions for the two approaches: $\eta = 0$ and $\eta = 1$. The average similarity score increases substantially from $\mathbf{0.162}$ when $\eta = 0$ to $\mathbf{0.350}$ when $\eta = 1$. We also adjusted and plotted two normal distributions $\mathcal{N}(\mu, \sigma)$ that fit the score distributions. A paired t-test yielded $t = -46.33$ and p-value $< 10^{-12}$, demonstrating that the proposed network δ considerably enhances feature merging, producing representations that more closely match the high-resolution gallery features.

4.3 General Results

In this section, we present the results obtained in our experiments. Tables 1 to 4 summarize the main findings. Dark gray highlights the best values, while light gray indicates the second-best.

Table 1 shows the quantitative results of our proposed method on the CelebA dataset. We used the gallery image as a reference for the first three metrics, Area Under the Curve (AUC), Rank-1, and Rank-5, and the high-resolution ground truth image as a reference for the remaining three metrics: PSNR, SSIM, and LPIPS. Compared to other algorithms, FASR++ provides superior results in all metrics, demonstrating its ability to recover discriminative identity features while maintaining high perceptual and structural image quality.

In general, the parameters of super-resolution algorithms are optimized to improve image quality metrics such as PSNR and SSIM; however, these improvements may be accompanied by distortions in identity and impaired recognition performance. Moreover, dos Santos *et al*. [2022] exemplifies that for a given set of solutions for a super-resolution problem, the images with higher values for PSNR and SSIM metrics are not the images that provide the highest value for recognition metrics. Remarkably, FASR++ maintains superior PSNR and SSIM values, and lower LPIPS scores, while still achieving optimal recognition performance by generating high-quality features through an ensemble of FC modules to assist the diffusion model.

Table 2 presents the quantitative results on the Quis-Campi dataset. The SRDG algorithm uses soft attributes as input to guide its diffusion process during image reconstruction. Although SRDG benefits from this additional conditioning, our method surpasses SRDG and the other competing algorithms in both recognition metrics (Rank-1 and Rank-5) and image quality metrics (PSNR, SSIM, and LPIPS).

To further assess the reliability of the image quality improvements, a paired t-test was conducted to evaluate the statistical significance of the results across both datasets. FASR++ was compared against the second-best performing methods: FASR for PSNR and SSIM, and SR3 for LPIPS. On the CelebA dataset, the improvements in PSNR and SSIM were confirmed to be statistically significant relative to FASR ($t_{\mathrm{PSNR}} = 28.9$, $t_{\mathrm{SSIM}} = 22.4$), while the difference in perceptual quality, measured by LPIPS, was also statistically significant compared to SR3 ($t_{\mathrm{LPIPS}} = -4.4$). Similarly, on the Quis-Campi dataset, statistically significant differences were observed when comparing FASR++ to FASR for PSNR and SSIM ($t_{\mathrm{PSNR}} = 26.4$, $t_{\mathrm{SSIM}} = 8.9$), and to SR3 for LPIPS ($t_{\mathrm{LPIPS}} = -16.1$). In all cases, the differences were highly significant ($p < 10^{-5}$), confirming that the observed gains are not due to random variation but instead establish the proposed method as the SOTA in both reconstruction fidelity and perceptual quality.

In Figures 6 and 7, we present the qualitative comparison of our method FASR++ against other super-resolution algorithms for the Celeba and Quis-Campi datasets, respectively. While all other methods are effective to some extent, they often introduce artifacts or noise into the facial images, typical

Table 1. Results for the CelebA Dataset. 1:1 verification and 1:N identification results, along with SSIM, PSNR, and LPIPS metrics, on both low-resolution (LR) and super-resolved versions.

SR Method	AUC	Rank-1 (%)	Rank-5 (%)	PSNR↑	SSIM↑	LPIPS↓
LR	0.885	11.80	29.60	23.2736 ± 1.9534	0.6452 ± 0.0711	0.4704 ± 0.0754
GFPGAN	0.865	19.80	34.80	23.1720 ± 1.7734	0.6545 ± 0.0688	0.1915 ± 0.0714
SPARNET	0.874	21.80	38.60	21.0168 ± 2.3292	0.6121 ± 0.0814	0.3014 ± 0.0869
SR3	0.936	44.60	62.60	25.4013 ± 1.9935	0.7370 ± 0.0677	0.1022 ± 0.0372
SwinIR	0.921	40.40	57.20	21.7619 ± 2.0440	0.6660 ± 0.0848	0.2008 ± 0.0702
SDE-SR	0.933	45.60	66.00	25.6987 ± 2.0236	0.7522 ± 0.0659	0.1116 ± 0.0392
FASR	0.946	53.20	68.60	25.8966 ± 2.0739	0.7588 ± 0.0657	0.1191 ± 0.0427
FASR++ (Ours)	0.951	59.60	74.80	26.6413 ± 2.1197	0.7756 ± 0.0643	0.0983 ± 0.0373

Table 2. Results for the Quis-Campi Dataset. 1:1 verification and 1:N identification results, along with SSIM, PSNR, and LPIPS metrics, on both low-resolution (LR) and super-resolved versions.

SR Method	AUC	Rank-1 (%)	Rank-5 (%)	PSNR↑	SSIM↑	LPIPS↓
LR	0.815	31.11	51.78	29.1295 ± 3.0910	0.8257 ± 0.0612	0.3150 ± 0.0814
GFPGAN	0.789	17.78	42.22	27.8496 ± 2.6545	0.7733 ± 0.0640	0.2575 ± 0.0557
SPARNET	0.862	32.67	58.44	24.9380 ± 3.9794	0.7663 ± 0.0986	0.2092 ± 0.0794
SR3	0.914	46.00	70.89	30.7066 ± 3.0262	0.8477 ± 0.0558	0.1235 ± 0.0372
SRDG	0.920	46.89	73.33	30.0598 ± 2.8980	0.8176 ± 0.0576	0.1410 ± 0.0459
IDM	0.884	30.89	59.11	26.2468 ± 3.6993	0.7522 ± 0.0866	0.1663 ± 0.0501
SDE-SR	0.916	47.11	71.56	30.3447 ± 2.8344	0.8250 ± 0.0546	0.1298 ± 0.0374
FASR	0.917	50.67	72.22	30.7092 ± 2.7705	0.8484 ± 0.0496	0.1303 ± 0.0342
FASR++ (Ours)	0.918	52.22	75.11	31.7817 ± 2.9570	0.8554 ± 0.0491	0.0983 ± 0.0309

issues encountered in SR algorithms. For instance, in most examples, the images generated by other algorithms exhibit distortions, mainly in the eye region, providing an artificial and distorted appearance. In Figure 6, GFPGAN yields distortions regarding the person's age. In contrast, FASR++ stands out as the only approach that produces natural-looking images without noticeable artificiality. It preserves symmetries and successfully recovers details without introducing artifacts or distorting facial features.

Due to the ill-posed nature of the SR problem, many SR algorithms suffer from bias issues and struggle to recover a person's identity accurately. In contrast, our algorithm effectively tackles these challenges, mitigating identity-related problems and yielding superior quantitative and qualitative results.

4.4 Ablation Study

This section analyzes the impact of the δ network and the choice of dataset on both recognition accuracy and image quality, followed by an evaluation of how the number of training samples used in the δ network influences its effectiveness and the overall performance of the model.

Influence of the δ Network and Dataset. In this section, we conduct an incremental study to examine the impact of each enhancement applied to our proposed method on the CelebA and Quis-Campi datasets. We start with **FASR**, where low-resolution features are combined using an arithmetic mean to produce a single super-resolution output. Next, we introduce **FASR++ ($\eta = 0$)**, which still combines low-resolution features via arithmetic mean but generates two super-resolution images that are averaged to create a final image. We also present two versions of our complete approach, **FASR++** and **FASR†++**. In both cases, an ensemble of Feature Combiner modules integrated with the δ network ($\eta = 1$) is employed to estimate a reliable descriptor and to produce two super-resolution images. As described in Section 4.2, in **FASR++** the δ network is trained on the CASIA-WebFace dataset using 10,572 identities, with an average of 46.41 images per subject. Alternatively, in **FASR†++**, the δ network is trained on a subset of the CelebA dataset, where we selected 2,000 identities with an average of 20 images per identity. This experiment aims to analyze the influence of dataset limitations such as fewer identities, reduced intra-class variability, and a smaller number of images per identity on the performance of the proposed fusion mechanism.

To further examine the role of the weighting factor η in the fusion process, we performed additional experiments with intermediate values $\eta \in \{0, 0.25, 0.5, 0.75, 1.0\}$ using the same experimental setup described in Section 4.2. For $\eta \neq 0$, the resulting cosine similarity distributions were largely overlapping, indicating that η has little effect on the fused representations. This behavior occurs because the FC module is trained to optimize the combined feature representation as a whole, so any scaling change in η tends to be internally compensated by the learned weights of the δ network, maintaining consistent FC outputs across different configurations. Therefore, $\eta = 1$ was adopted in all experiments.

Tables 3 and 4 present the verification and recognition results along with the corresponding image quality metrics.

Figure 6. Qualitative results for the CelebA Dataset. Comparison of low-resolution (LR) images, super-resolution (SR) outputs obtained by various methods, and ground truth (GT) images. FASR++ outperforms the baselines by preserving facial symmetry and ensuring a natural appearance.

In addition, the detailed CMC curves for the CelebA and Quis-Campi datasets are shown in Figure 8. From Table 3, we observe a consistent improvement as each component is introduced in our incremental study. Averaging two images in FASR++ ($\eta = 0$) slightly enhances reconstruction quality compared to the baseline FASR, while maintaining similar recognition performance. When the learned fusion mechanism is activated ($\eta = 1$) in FASR†++, both recognition accuracy and perceptual quality improve, with a more pronounced effect on recognition (a 4% increase in Rank-1 accuracy). Finally, training δ on the larger CASIA-WebFace dataset (FASR++) yields the best overall Rank-1 accuracy (a 6.4% improvement over the baseline), whereas for the other metrics, both FASR†++ and FASR++ achieve very similar results.

From Table 4, we observe a similar incremental behavior on the Quis-Campi dataset. Averaging two images in FASR++ ($\eta = 0$) results in a clear perceptual improvement, yielding a substantial reduction in LPIPS (from 0.1303 to 0.1004) compared to FASR. When the learned fusion mechanism is enabled ($\eta = 1$) in FASR†++, the gain is more pronounced in Rank-5 accuracy. Finally, training δ on the larger CASIA-WebFace dataset (FASR++) yields the best overall results, except for Rank-5 accuracy.

For both datasets, we observe that using the δ network ($\eta = 1$) provides superior results compared to the other configurations. This outcome was expected since, as discussed in Section 4.2, the δ network combines features efficiently. The improved results in both recognition and image quality metrics confirm that the combined features are being effectively leveraged by the diffusion model. Moreover, the larger number of images and the greater identity diversity of the CASIA-WebFace dataset contribute to the superior results of FASR++ compared to FASR†++, mainly in terms of Rank-1 accuracy.

We can also observe that the superior results of FASR++ are more pronounced for the CelebA dataset and more subtle for Quis-Campi. This can be attributed to two main reasons: (i) as will be described in the next section, the Quis-Campi dataset is a real-world surveillance dataset, which is more challenging, containing variations in lighting and pose, as well as a higher level of noise compared to CelebA. Consequently, the facial features in Quis-Campi can be noisier and more difficult to reconstruct accurately; (ii) for the Quis-Campi dataset, only five features from low-resolution images were used to assist in the reconstruction of the super-resolution images, while more than eight low-resolution images were used for CelebA, providing a richer source of information for the algorithm.

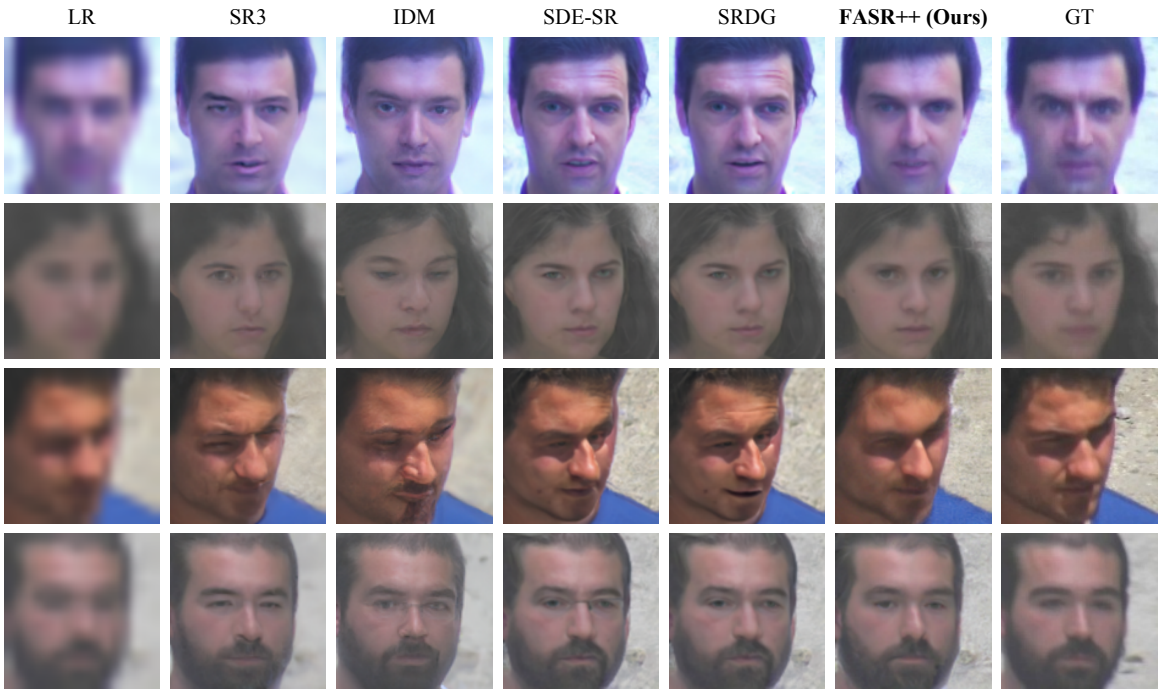

Figure 7. Qualitative Results for the Quis-Campi Dataset. Comparison of low-resolution (LR) images, super-resolution (SR) outputs obtained by various methods, and ground truth (GT) images. FASR++ outperforms the baselines by preserving facial symmetry and ensuring a natural appearance.

Table 3. Ablation Study on the CelebA dataset. We evaluate the impact of our FASR++ approach by comparing three configurations: FASR, FASR++ with $\eta = 0$, and FASR†++ with $\eta = 1$. In the latter case, the δ network was trained on a subset of the CelebA dataset.

SR Method	η	AUC	Rank-1 (%)	Rank-5 (%)	PSNR↑	SSIM↑	LPIPS↓
FASR	0	0.946	53.20	68.60	25.8966 ± 2.0739	0.7588 ± 0.0657	0.1191 ± 0.0427
FASR++	0	0.944	53.20	68.80	26.5421 ± 2.0690	0.7701 ± 0.0637	0.0989 ± 0.0362
FASR†++	1	0.950	57.20	75.20	26.5531 ± 2.0815	0.7719 ± 0.0641	0.0938 ± 0.0351
FASR++	1	0.951	59.60	74.80	26.6413 ± 2.1197	0.7756 ± 0.0643	0.0983 ± 0.0373

Table 4. Ablation Study on the Quis-Campi dataset. We evaluate the impact of our FASR++ approach by comparing three configurations: FASR, FASR++ with $\eta = 0$, and FASR†++ with $\eta = 1$. In the latter case, the δ network was trained on a subset of the CelebA dataset.

SR Method	η	AUC	Rank-1 (%)	Rank-5 (%)	PSNR↑	SSIM↑	LPIPS↓
FASR	0	0.917	50.67	72.22	30.7092 ± 2.7705	0.8484 ± 0.0496	0.1303 ± 0.0342
FASR++	0	0.915	48.44	73.78	31.7033 ± 2.9282	0.8531 ± 0.0487	0.1004 ± 0.0310
FASR†++	1	0.918	50.22	75.56	31.6540 ± 2.9476	0.8516 ± 0.0511	0.1011 ± 0.0319
FASR++	1	0.918	52.22	75.11	31.7817 ± 2.9570	0.8554 ± 0.0491	0.0983 ± 0.0309

Effect of the Training Dataset Size on the δ Network. We now analyze how the total number of samples used for training the δ network affects its ability to effectively merge features and influence the final results of the FASR++ framework, using the CelebA dataset as the evaluation benchmark. In this experiment, we limit the number of images per identity in the CASIA-WebFace dataset and evaluate the impact on the PSNR, SSIM, Rank-1, and Rank-5 metrics. Figure 9 shows that as the number of training samples increases, both image quality metrics (PSNR and SSIM) and recognition metrics (Rank-1 and Rank-5 accuracies) remain relatively stable with only minor fluctuations, suggesting that approximately 50k images are sufficient to achieve optimal performance in both reconstruction quality and face recognition. In this range, the Rank-1 and Rank-5 scores converge to around 60% and 75%, respectively, while statistical analysis confirmed that, despite small variations in PSNR and SSIM, no significant differences were observed when training with 50k images or more ($p > 0.05$).

4.5 Failure Cases

Figure 10 shows some failure cases of our algorithm compared to SRDG and SDE-SR. In the first row, FASR++ fails to recover the eyeglasses correctly, whereas SRDG successfully recovers this attribute. However, it is important to note that SRDG requires explicit information on whether the person is wearing eyeglasses. This information is not always discernible from LR images in surveillance scenarios.

In the second row of Figure 10, we observe a failure case

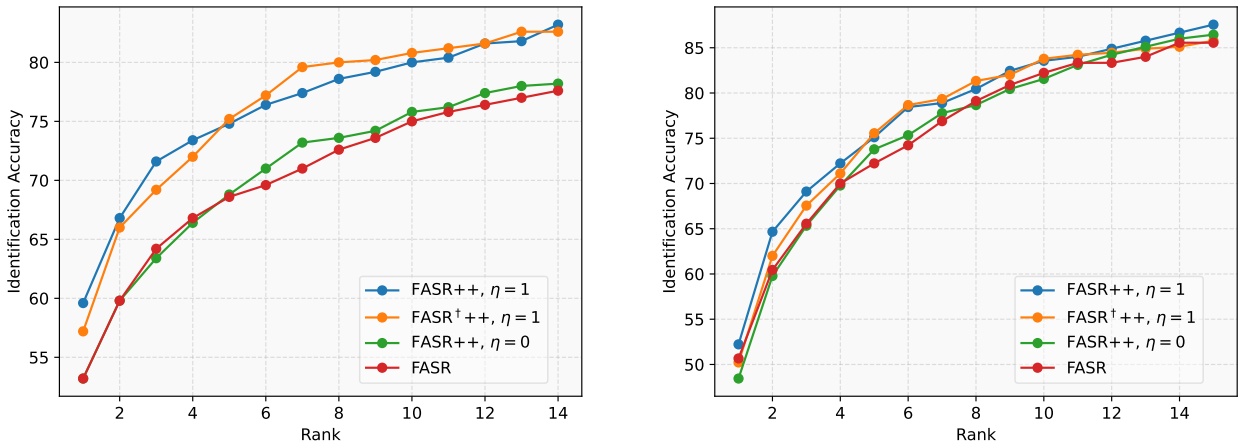

Figure 8. CMC curves on CelebA (left) and Quis-Campi (right) datasets. The evaluation includes four super-resolution methods: FASR, FASR++ with $\eta = 0$ and $\eta = 1$, and FASR†++ (where the δ network was trained on a subset of the CelebA dataset). The proposed FASR++ improves recognition accuracy across both datasets, demonstrating its effectiveness over the baseline methods.

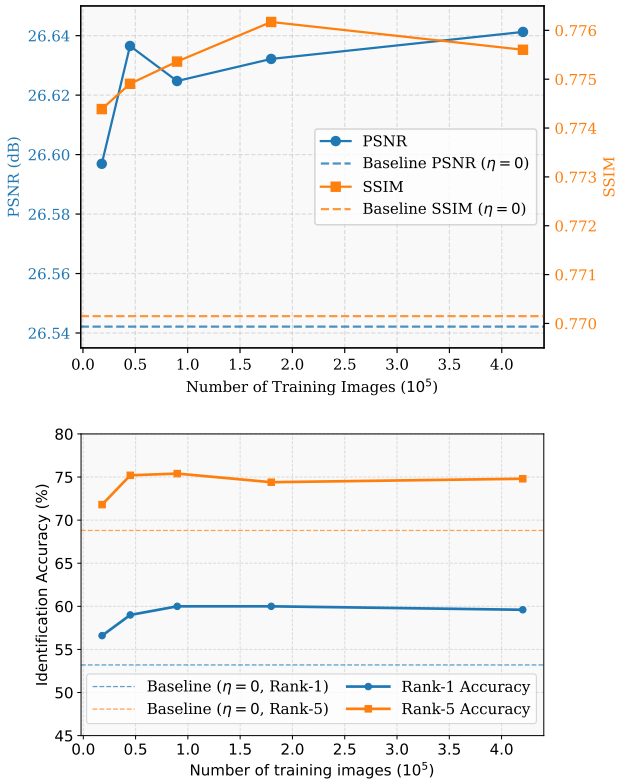

Figure 9. Performance metrics of FASR++ as a function of the number of training images. The top plot illustrates the image quality metrics (PSNR and SSIM), whereas the bottom plot presents the recognition metrics (Rank-1 and Rank-5 accuracies).

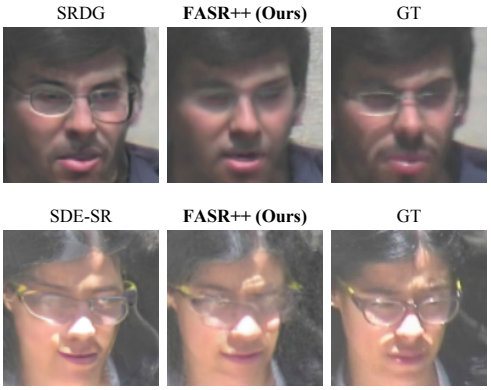

Figure 10. Failure cases. The first row presents results from SRDG [dos Santos *et al.*, 2024b], FASR++ (ours), and ground truth (GT) images, while the second row presents results from SDE-SR [dos Santos *et al.*, 2022], FASR++ (ours), and GT images.

of FASR++ compared to SDE-SR. The image in question shows significant pose variation and highly heterogeneous illumination. FASR++ produces smoother images with less noise than the other algorithms, causing the information about eyeglasses and the sun's reflection to spread across the periocular region.

Upon closer examination of the cases where our algorithm fails in Rank-5, we observed that most images share characteristics similar to those described in the previous paragraphs. Thus, FASR++ provides better results for recognition accuracy but may be more sensitive to variations in pose and lighting.

5 Conclusions

In this work, we introduced FASR++, an algorithm that effectively combines multiple features through a neural network to produce a reliable and representative feature vector. This vector is then integrated with a reference low-resolution image in a diffusion model to generate high-quality super-resolution images. A key advantage of our algorithm is its independence from explicitly provided facial attributes; instead, it implicitly extracts high-level information through a visual encoder. This methodology enables our algorithm to preserve individuals' identities more effectively than other methods, resulting in high-quality SR images with enhanced face symmetry, reduced noise and minimized distortion of facial attributes. We validated our approach on the CelebA and Quis-Campi datasets and achieved state-of-the-art results for visual quality and recognition metrics, demonstrating its potential for applications in real-world surveillance scenarios.

Declarations

Authors' Contributions

Marcelo dos Santos is the main contributor and writer of this manuscript. Rayson Laroca assisted in reviewing and editing the manuscript. João Carlos Raposo Neves co-supervised the project and reviewed the manuscript. David Menotti supervised the project, provided resources, and contributed to the review and editing. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was financed by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES)* – Finance Code001, by the "V2IP: Videomonitoramento para Identificação de Pessoas e Veículos" CAPES-PROCAD project (# 88887.619562/2021-00), and by the *Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)* (# 315409/2023-1).

Availability of data and materials

Our code is publicly available at `https://github.com/marcelowds/fasrpp`.

References

Anderson, B. D. (1982). Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326. DOI: 10.1016/0304-4149(82)90051-5.

Baker, S. and Kanade, T. (2002). Limits on super-resolution and how to break them. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1167–1183. DOI: 10.1109/TPAMI.2002.1033210.

Bilgazyev, E., Efraty, B., Shah, S. K., and Kakadiaris, I. A. (2011). Improved face recognition using super-resolution. In *International Joint Conference on Biometrics (IJCB)*, pages 1–7. DOI: 10.1109/IJCB.2011.6117554.

Cai, R., Yang, G., Averbuch-Elor, H., Hao, Z., Belongie, S., Snavely, N., and Hariharan, B. (2020). Learning gradient fields for shape generation. In *European Conference on Computer Vision (ECCV)*, pages 364–381. DOI: 10.1007/978-3-030-58580-8_22.

Chen, C., Gong, D., Wang, H., Li, Z., and Wong, K.-Y. K. (2020). Learning spatial attention for face super-resolution. *IEEE Transactions on Image Processing*, 30:1219–1231. DOI: 10.1109/TIP.2020.3043093.

dos Santos, M., Laroca, R., Ribeiro, R. O., Neves, J., and Menotti, D. (2024a). Multi-feature aggregation in diffusion models for enhanced face super-resolution. In *Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 1–6. DOI: 10.1109/SIBGRAPI62404.2024.10716316.

dos Santos, M., Laroca, R., Ribeiro, R. O., Neves, J., Proença, H., and Menotti, D. (2022). Face super-resolution using stochastic differential equations. In *Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 216–221. DOI: 10.1109/SIBGRAPI55357.2022.9991799.

dos Santos, M., Neves, J. C. R., Proença, H., and Menotti, D. (2024b). Defying limits: Super-resolution refinement with diffusion guidance. In *International Conference on Computer Vision Theory and Applications (VISAPP)*, pages 426–434. DOI: 10.5220/0012398900003660.

Gao, S., Liu, X., Zeng, B., Xu, S., Li, Y., Luo, X., Liu, J., Zhen, X., and Zhang, B. (2023). Implicit diffusion models for continuous super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10021–10030. DOI: 10.1109/CVPR52729.2023.00966.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. DOI: 10.1109/CVPR.2016.90.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851. DOI: 10.5555/3495724.3496298.

Jiang, J., Wang, C., Liu, X., and Ma, J. (2021). Deep learning-based face super-resolution: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36. DOI: 10.1145/3485132.

Jolicoeur-Martineau, A., Li, K., Piché-Taillefer, R., Kachman, T., and Mitliagkas, I. (2021). Gotta go fast when generating data with score-based models. *arXiv preprint*. DOI: 10.48550/arXiv.2105.14080.

Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405. DOI: 10.1109/CVPR.2019.00453.

Kim, M., Jain, A. K., and Liu, X. (2022). AdaFace: Quality adaptive margin for face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI: 10.1109/CVPR52688.2022.01201.

Kloeden, P. and Platen, E. (2011). *Numerical Solution of Stochastic Differential Equations*, volume 23. Springer. DOI: 10.1007/978-3-662-12616-5.

Labach, A., Salehinejad, H., and Valaee, S. (2019). Survey of dropout methods for deep neural networks. *arXiv preprint*. DOI: 10.48550/arXiv.1904.13310.

Lee, C.-H., Zhang, K., Lee, H.-C., Cheng, C.-W., and Hsu, W. (2018). Attribute augmented convolutional neural network for face hallucination. In *IEEE Conference on Computer Vision and Pattern Recognition workshops*, pages 721–729. DOI: 10.1109/CVPRW.2018.00115.

Li, H., Yang, Y., Chang, M., Chen, S., Feng, H., Xu, Z., Li, Q., and Chen, Y. (2022). SRDiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59. DOI: 10.1016/j.neucom.2022.01.029.

Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., and Timofte, R. (2021). Swinir: Image restoration using swin transformer. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1833–1844. DOI: 10.1109/ICCVW54120.2021.00210.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *IEEE International Confer-*

ence on Computer Vision (ICCV), pages 3730–3738. DOI: 10.1109/ICCV.2015.425.

Lu, Y., Tai, Y.-W., and Tang, C.-K. (2018). Attribute-guided face generation using conditional cyclegan. In *European Conference on Computer Vision (ECCV)*, pages 282–297. DOI: 10.1007/978-3-030-01258-8_18.

Meng, C., Rombach, R., Gao, R., Kingma, D., Ermon, S., Ho, J., and Salimans, T. (2023). On distillation of guided diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14297–14306. DOI: 10.1109/CVPR52729.2023.01374.

Nascimento, V., Laroca, R., Lambert, J. A., Schwartz, W. R., and Menotti, D. (2022). Combining attention module and pixel shuffle for license plate super-resolution. In *Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 228–233. DOI: 10.1109/SIBGRAPI55357.2022.9991753.

Nascimento, V., Laroca, R., Ribeiro, R. O., Schwartz, W. R., and Menotti, D. (2024). Enhancing license plate super-resolution: A layout-aware and character-driven approach. *Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 1–6. DOI: 10.1109/SIBGRAPI62404.2024.10716303.

Neves, J., Moreno, J., and Proença, H. (2018). QUIS-CAMPI: an annotated multi-biometrics data feed from surveillance scenarios. *IET Biometrics*, 7(4):371–379. DOI: 10.1049/iet-bmt.2016.0178.

Niu, C., Song, Y., Song, J., Zhao, S., Grover, A., and Ermon, S. (2020). Permutation invariant graph generation via score-based generative modeling. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108, pages 4474–4484. DOI: 10.48550/arXiv.2003.00638.

Richter, J., Frintrop, S., and Gerkmann, T. (2023). Audio-visual speech enhancement with score-based generative models. In *ITG Conference on Speech Communication*, pages 275–279. DOI: 10.48550/arXiv.2306.01432.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. DOI: 10.1007/978-3-319-24574-4$_2$8.

Saharia, C., Chan, W., Saxena, S., Lit, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Gontijo-Lopes, R., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 36479–36494. DOI: 10.5555/3600270.3602913.

Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., and Norouzi, M. (2023). Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726. DOI: 10.1109/TPAMI.2022.3204461.

Särkkä, S. and Solin, A. (2019). *Applied stochastic differential equations*, volume 10. Cambridge University Press. Book.. DOI: 10.1017/9781108186735.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, pages 2256–2265. DOI: 10.48550/arXiv.1503.03585.

Song, Y. and Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1–13. DOI: 10.5555/3454287.3455354.

Song, Y. and Ermon, S. (2020). Improved techniques for training score-based generative models. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:12438–12448. DOI: 10.5555/3495724.3496767.

Song, Y., Shen, L., Xing, L., and Ermon, S. (2022). Solving inverse problems in medical imaging with score-based generative models. In *International Conference on Learning Representations (ICLR)*, pages 1–18. DOI: 10.48550/arXiv.2111.08005.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2021). Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, pages 1–36. DOI: 10.48550/arXiv.2011.13456.

Suin, M., Nair, N. G., Pong Lau, C., Patel, V. M., and Chellappa, R. (2024). Diffuse and restore: A region-adaptive diffusion model for identity-preserving blind face restoration. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6343–6352. DOI: 10.1109/WACV57701.2024.00622.

Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., and Ng, R. (2020). Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:7537–7547. DOI: 10.5555/3495724.3496356.

Vahdat, A., Kreis, K., and Kautz, J. (2021). Score-based generative modeling in latent space. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 11287–11302. DOI: 10.5555/3540261.3541124.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30. DOI: 10.5555/3295222.3295349.

Vincent, P. (2011). A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674. DOI: 10.1162/NECO_a_00142.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., and Tang, X. (2017). Residual attention network for image classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164. DOI: 10.1109/CVPR.2017.683.

Wang, X., Li, Y., Zhang, H., and Shan, Y. (2021). Towards real-world blind face restoration with generative facial prior. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9164–9174. DOI: 10.1109/CVPR46437.2021.00905.

Yi, D., Lei, Z., Liao, S., and Li, S. Z. (2014). Learning face representation from scratch. *arXiv preprint*. DOI: 10.48550/arXiv.1411.7923.

Yu, X., Fernando, B., Hartley, R., and Porikli, F. (2018).

Super-resolving very low-resolution face images with supplementary attributes. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 908–917. DOI: 10.1109/CVPR.2018.00101.

Yuan, Y., Chen, W., Yang, Y., and Wang, Z. (2020). In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1454–1463. DOI: 10.1109/CVPRW50498.2020.00185.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595. DOI: 10.1109/CVPR.2018.00068.

Zhang, Z., Han, L., Ghosh, A., Metaxas, D., and Ren, J. (2023). Sine: Single image editing with text-to-image diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6027–6037. DOI: 10.1109/CVPR52729.2023.00584.

Zhu, S., Liu, S., Loy, C. C., and Tang, X. (2016). Deep cascaded bi-network for face hallucination. In *European Conference on Computer Vision (ECCV)*, pages 614–630. DOI: 10.1007/978-3-319-46454-1$_3$7.