Multi-Feature Aggregation in Diffusion Models for Enhanced Face Super-Resolution

Marcelo dos Santos*, Rayson Laroca†,*, Rafael O. Ribeiro‡, João C. Neves§, David Menotti*

*Federal University of Paraná, Curitiba, Brazil †Pontifical Catholic University of Paraná, Curitiba, Brazil
‡Brazilian Federal Police, Brasília, Brazil §University of Beira Interior, Covilhã, Portugal

*{msantos,menotti}@inf.ufpr.br †rayson@ppgia.pucpr.br ‡rafael.ror@pf.gov.br §jcneves@di.ubi.pt

***Abstract*—Super-resolution algorithms often struggle with images from surveillance environments due to adverse conditions such as unknown degradation, variations in pose, irregular illumination, and occlusions. However, acquiring multiple images, even of low quality, is possible with surveillance cameras. In this work, we develop an algorithm based on diffusion models that utilize a low-resolution image combined with features extracted from multiple low-quality images to generate a super-resolved image while minimizing distortions in the individual's identity.**

Unlike other algorithms, our approach recovers facial features without explicitly providing attribute information or without the need to calculate a gradient of a function during the reconstruction process. To the best of our knowledge, this is the first time multi-features combined with low-resolution images are used as conditioners to generate more reliable super-resolution images using stochastic differential equations. The FFHQ dataset was employed for training, resulting in state-of-the-art performance in facial recognition and verification metrics when evaluated on the CelebA and Quis-Campi datasets. Our code is publicly available at *https://github.com/marcelowds/fasr*.

I. Introduction

The problem of super-resolution (SR) is inherently ill-posed, making the recovery of fine details like eyeglasses, beards, mustaches, and a reliable identity quite challenging [1], [2]. For surveillance scenarios, the presence of noise, occlusions, variations in illumination, and varying poses make the problem even harder, leading to a significant decline in the performance of SR and face recognition algorithms [3], [4].

Soft biometrics, such as gender, hair color, and skin tone, can enhance image reconstruction, reducing the ambiguity in face SR and increasing the reliability of recognition systems [5]. However, facial attributes are often not visible in low-resolution (LR) images, making reliable access challenging. Also, obtaining these attributes typically requires a classifier or manual extraction, which is not very efficient [6].

In [5] and [6], attributes such as beard and glasses, among others, are used to improve the quality of SR algorithms. Nevertheless, these attributes alone are insufficient to generate accurate high-resolution images. It is also necessary to consider subtle characteristics, such as facial proportions, shapes, and other high-level, more abstract features. Therefore, it is essential to develop algorithms that rely on more general characteristics as sources of information for image reconstruction.

Diffusion models are used for data generation across diverse domains, and here, they are employed to generate SR images. These models operate by adding noise at different scales to the data and training a network to predict the noise present. Once trained, the network can perform reverse diffusion, removing the noise and generating the desired type of data.

Diffusion models can integrate various types of information, such as text and image [7], image and audio [8], and multi-modal data [9]. This enables the generation of extraordinarily high-quality and original data. This concept is central to our work, where we combine LR images with facial features.

Another tool commonly used in conjunction with diffusion models is the classifier guidance method, which is used to generate data within predefined classes or with specific characteristics. It involves utilizing a classifier's gradient as a supervisor during reverse diffusion. For instance, it was used in [6] to recover facial attributes. A drawback of this strategy is the need to train a classifier, along with the additional computational cost associated with gradient calculations.

The main contribution of our work lies in developing Feature Aggregation Super-Resolution (FASR), an SR algorithm that recovers crucial features for face recognition. In addition to the LR image, it takes as input a vector of facial features derived from a set of LR images, which can be either a series of video frames or independent images of an individual. This new vector has a higher signal-to-noise ratio than each vector individually. It is incorporated into the network, merging its information with LR image to generate an SR version. In this way, our algorithm effectively recovers facial information from an image, yielding results of higher quality with minimal distortion of identity. FASR employs diffusion models based on a Stochastic Differential Equation (SDE) and operates without the need for a classifier to guide the reverse diffusion process.

Our method's effectiveness has been validated on the CelebA and QuisCampi datasets. Our SR algorithm produces superior qualitative and quantitative results. The state-of-the-art values are supported by better results in face recognition metrics such as Area Under the Curve (AUC) in the 1:1 verification protocol and accuracy in the 1:N identification protocol.

This paper is structured as follows. Section II outlines related works, Section III describes the main concept behind the proposed method, and Section IV details our experiments and results. Finally, Section V concludes the paper.

Accepted at SIBGRAPI 2024. The final published version is available on IEEE Xplore (DOI: 10.1109/SIBGRAPI62404.2024.10716316).

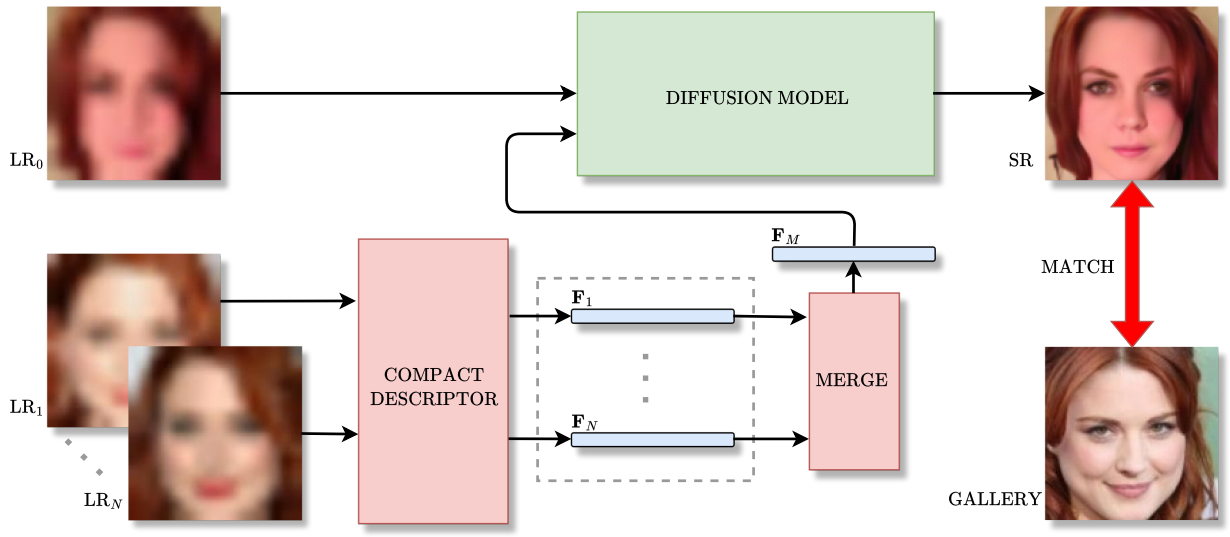

Fig. 1. Overview of the proposed method. The low-resolution images $\mathrm{LR}_1, \ldots, \mathrm{LR}_N$ are used to compute a set of features $\mathbf{F}_1, \ldots, \mathbf{F}_N$, respectively, which are then combined to generate $\mathbf{F}_M$. The low-resolution image LR_0 is integrated with $\mathbf{F}_M$ in the diffusion model to produce a super-resolution (SR) image. The SR image is subsequently compared with a set of images from the gallery for face recognition.

II. Related Work

In the seminal work [10], Sohl-Dickstein et al. utilized principles from non-equilibrium thermodynamics to create a generative model. Two other works in the line of diffusion models that had much impact in this field are Denoising Diffusion Probabilistic Models (DDPMs) [11] and Score-Based Generative Models (SGMs) [12], [13]. In [14], DDPM and SGD are generalized for continuous time steps and noise levels using Stochastic Differential Equations (SDEs), expanding the range of possibilities of research 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 [11], [12], [14]–[17]. The full potential of diffusion models can also be leveraged through multi-domain data integration, such as text-to-image translation [7] and image editing [18]. Additionally, [8] combines audio-visual information for speech enhancement.

SR is another significant application of diffusion models, which is utilized in this work. In [19], an adaptation of the DDPM model produces high-quality SR images. Similarly, SRDiff [20] employs diffusion models to estimate the difference between the original LR image and an high-resolution (HR) image, resulting in an SR image. In [21], SDEs were used to generate SR images. Additionally, [6] performs SR by incorporating attribute information such as beard, gender, and the presence of eyeglasses to generate high-quality images. However, this approach has the drawback that these attributes must be explicitly provided to the algorithm, which cannot be easily estimated in LR images.

In [22], an identity-preserving SR method was developed. In both [6] and [22], 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 excellent results achieved by diffusion models, a major drawback is their high execution time due to their iterative nature. However, this issue is likely to be mitigated in the mid-term, as several studies are focused on enhancing the computational efficiency of these methods. For a more detailed discussion on accelerating and improving the efficiency of sampling in diffusion models, refer to [23]–[25].

III. Proposed Method

In this section, we present the general idea of our proposed method, followed by a brief theoretical background on diffusion models based on SDEs and a description of how the facial features are incorporated into the model.

A. General Idea

As previously noted, images captured in surveillance environments are often low-quality. Nevertheless, in certain instances, a video of a particular person can provide multiple LR images that can be combined to enhance the performance of SR algorithms. This combination of information from multiple images is expected to increase the signal-to-noise ratio, providing higher-quality information.

In this study, we aim to improve the performance of an SR algorithm by integrating a reference LR image with a combination of multiple features extracted from different LR images (see Fig. 1). This integration leads to enhanced image 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.

B. 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 [14], [26], a forward diffusion process $\{\mathbf{x}(t)\}_{t=0}^{T}$ and its reverse are, respectively, modeled using the following SDEs:

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}, \tag{1}$$

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

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 former runs backward in time) and p_t is the probability density of $\mathbf{x}(t)$. References [27], [28] 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 an SR image, we need to solve Equation 2, and for that, we use a deep neural network s_θ to approximate $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$. The neural network is conditioned on LR images and image features, denoted by $\mathbf{y}$ and $\mathbf{F}_M$, respectively. The training of the neural network $s_\theta(\mathbf{x}(t), \mathbf{y}, \mathbf{F}_M, t)$ is achieved by optimizing the following loss function [29]:

$$\min_{\theta} \mathbb{E}_{t\sim\mathcal{U}[0,T]} \mathbb{E}_{\mathbf{x}_0\sim p(\mathbf{x}_0)} \mathbb{E}_{\mathbf{x}(t)\sim p_t(\mathbf{x}(t)|\mathbf{x}(0))} \big[\lambda(t) \times \|s_\theta(\mathbf{x}(t), \mathbf{y}, \mathbf{F}_M, t) - \nabla_{\mathbf{x}(t)} \log p(\mathbf{x}(t)|\mathbf{x}(0))\|_2^2\big], \tag{3}$$

where $\lambda(t)$ is a positive weighting function and $p(\mathbf{x}(t)|\mathbf{x}(0))$ is the transition kernel from $\mathbf{x}(0)$ to $\mathbf{x}(t)$.

Here, we use the variation exploding (VE) case described in [14] with $\mathbf{f}(\mathbf{x}, t)$ and $g(t)$ given respectively by:

$$\mathbf{f}(\mathbf{x}, t) = \mathbf{0}, \quad g(t) = \sqrt{\frac{\mathrm{d}\sigma^2(t)}{\mathrm{d}t}}, \tag{4}$$

where $\sigma(t) = \sigma_{min} \left(\sigma_{max}/\sigma_{min}\right)^t$ 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 [14]:

$$\boldsymbol{\mu}(t) = \mathbf{x}(0), \ \boldsymbol{\Sigma}(t) = [\sigma^2(t) - \sigma^2(0)]\mathbf{I}. \tag{5}$$

Thus, we can analytically compute $\nabla_{\mathbf{x}} \log p(\mathbf{x}(t)|\mathbf{x}(0))$ in Equation 3, allowing for efficient model training. Once the network well estimates the gradient, we change $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ by $s_\theta(\mathbf{x}(t), t)$ in the reverse process (Equation 2) and solve it from $t = T$ to $t = 0$ using the Euler-Maruyama method [27], [28] to generate an SR image $\mathbf{x}(0)$.

C. Model Conditioning

As in most diffusion models, we employ the U-Net architecture [11]. To condition the model on LR images, we follow a method similar to the one outlined in [19], [21]. This method involves concatenating in the channel domain, the LR image $\mathbf{y}$ and $\mathbf{x}_T$, which is the image undergoing denoising. This concatenation results in a 6-channel image.

The network is conditioned on image features using a method similar to time and class conditioning described in [30]. For a given level of the layers, the weighted sum of the time embedding and the feature vector is added to the image, providing a conditioned image, as shown in Fig. 2.

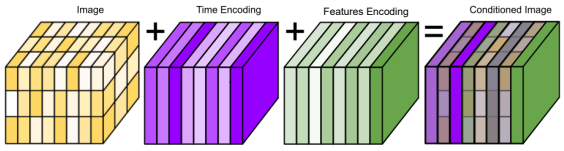

Fig. 2. Time and features encoding. Adapted from [31].

We used a compact descriptor (see Section IV-A for more details) extracted from the HR images during training to condition the neural network. The LR images are upsampled to the original size of HR images to preserve the dimensions. During the evaluation phase, one LR_0 image is used as a reference (in general, one should select the best and most frontal image as the LR_0 reference image, but here it was chosen randomly), while other images $\mathrm{LR}_1, \ldots, \mathrm{LR}_N$ are used to compute feature vectors $\mathbf{F}_1, \ldots, \mathbf{F}_N$. The merging of feature vectors can be performed in various ways. In this work, the merged $\mathbf{F}_M$ is the arithmetic mean of all feature vectors $\mathbf{F}_1, \ldots, \mathbf{F}_N$. We are assuming that $\mathbf{F}_M$ has a higher signal-to-noise ratio than a feature vector obtained from an individual LR image and that this mean vector approximately represents the actual characteristics of the HR image.

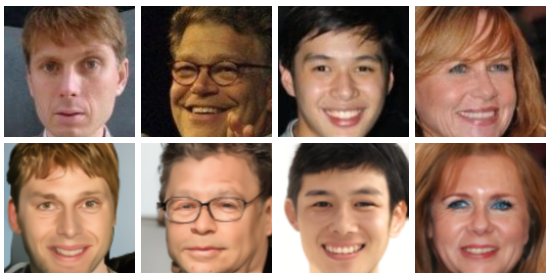

Fig. 3. First row: original HR images extracted from the CelebA dataset. Second row: synthetic HR images generated using only the feature vector extracted from the corresponding image above.

To demonstrate the efficacy of utilizing the feature vector to generate an SR image, we trained Equation 3 with $\mathbf{y} = 0$ and exclusively employed the feature vector to produce some images, as depicted in Fig. 3. These images showcase the algorithm's ability to reconstruct crucial high-level features necessary for preserving a person's identity. When coupled with the LR image, the feature vector proves effective in restoring image details while minimizing identity distortion.

IV. Experiments and Results

This section describes the experimental setup, followed by the results obtained on two distinct datasets. Lastly, we examine some extreme cases where the algorithm may fail.

A. Experiments

In this study, we explored three datasets: FFHQ [32], CelebA [33], and Quis-Campi [34], all gathered from surveillance scenarios. The FFHQ dataset was employed for model training, where 10^6 training steps were conducted. CelebA was employed 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 a LR probe image. The remaining images were also downsampled and used to extract features, assisting the reconstruction process 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. 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 generate the SR images. 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 training the SR algorithm and for facial recognition consists of a 512-dimensional vector generated through AdaFace [35] with a ResNet backbone [36] trained on the CASIA-WebFace dataset [37]. 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 state-of-the-art algorithms: SR3 [19] and SDE-SR [21].

B. Results

Table I shows the quantitative results of our algorithm on the CelebA dataset. Notably, FASR provides superior performance in terms of AUC, Rank-1, Rank-5, and Rank-10 (with an improvement of up to 4%) compared to other algorithms. As our algorithm incorporates the feature vector during the image generation phase, the images can be recovered while maintaining features that positively impact recognition and verification. Table II shows the quantitative results on the Quis-Campi dataset. In this context, we made additional comparisons with IDM [38] and SRGD [6], state-of-the-art algorithms. IDM represents an enhancement of SR3, while SRDG takes feature information as input and attempts to incorporate these features in SR images. Our algorithm provides superior results in recognition accuracy (Rank-1).

TABLE I
THE 1:1 VERIFICATION AND 1:N IDENTIFICATION (RANK-1, RANK-5 AND RANK-10) RESULTS OBTAINED USING THE ADAFACE RECOGNITION MODEL THROUGH SUPER-RESOLUTION ON THE CELEBA DATASET.

SR Method	AUC	Rank-1 (%)	Rank-5 (%)	Rank-10 (%)
LR	0.885	27.00	41.40	51.60
SR3	0.936	45.60	62.00	71.00
SDE-SR	0.933	48.60	66.60	72.40
FASR (Ours)	0.946	52.80	70.00	76.00

TABLE II
THE 1:1 VERIFICATION AND 1:N IDENTIFICATION (RANK-1, RANK-5 AND RANK-10) RESULTS OBTAINED USING THE ADAFACE RECOGNITION MODEL THROUGH SUPER-RESOLUTION ON THE QUIS-CAMPI DATASET.

SR Method	AUC	Rank-1(%)	Rank-5(%)	Rank-10(%)
LR	0.816	23.78	46.89	58.67
IDM	0.885	28.22	56.44	70.00
SR3	0.914	45.78	69.56	79.77
SDE-SR	0.917	50.00	72.67	81.56
SRDG	0.920	49.33	73.11	82.00
FASR (Ours)	0.917	51.33	72.44	80.00

Lastly, the qualitative outcomes for on the Quis-Campi dataset are presented in Fig. 4. Our method, FASR, is compared against other methods utilizing diffusion models to restore facial details and features. While these methods are effective to some extent, they often introduce artifacts or noise onto the facial images, typical issues encountered in SR algorithms. In contrast, FASR stands out as the only approach that produces natural-looking images without noticeable artificiality. It maintains facial naturalness, preserves symmetries, and successfully recovers details without introducing artifacts or distorting facial features.

For instance, in row 3 of Fig. 4, images generated by other algorithms exhibit distortions, particularly in the eye region, leading to a loss of naturalness and symmetry in the faces. Conversely, in row 4 of Fig. 4, our algorithm produces images with significantly reduced noise compared to the others. Moving to row 6 of Fig. 4, images generated by the SDE-SR and SRDG algorithms appear to "age" the subject, whereas our algorithm preserves the person's age while maintaining their identity without distortion.

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. Our algorithm tackles this by fusing a reference LR image with a multi-feature vector, effectively mitigating identity-related problems and yielding superior quantitative and qualitative results. However, additional tests and experiments are required to reduce bias and identity distortions before deploying this algorithm in real-world scenarios, especially in surveillance environments characterized by noisy and more challenging data, where errors in facial recognition can have adverse consequences.

C. Failure Cases

Fig. 5 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

Fig. 4. Comparison of low-resolution (LR), super-resolution (SR) results obtained by various methods, and ground gruth (GT) images from the Quis-Campi dataset. FASR outperforms baseline methods, preserving facial symmetry and natural appearance.

this attribute. Nevertheless, it is important to note that SRDG requires explicit information about whether the person is wearing eyeglasses. This information is not always discernible from LR images in surveillance scenarios.

In the second row of Fig. 5, we observe a failure case 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.

V. Conclusions

In this work, we introduced FASR, an algorithm that integrates multi-features and a reference LR image into diffusion models to generate SR images. A key advantage of our algorithm is its independence from explicitly provided facial attributes; instead, it utilizes features extracted using a deep neural network. 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 distortions in face attributes. Our approach was validated on the CelebA and Quis-Campi datasets, where we achieved state-of-the-art results for recognition metrics. Hence, it demonstrates the potential to be applied in real-world surveillance scenarios.

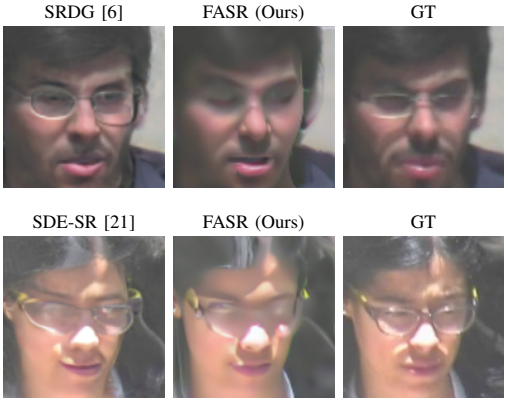

Fig. 5. Failure cases: the first row presents results from SRDG, FASR (ours), and ground truth (GT) images, while the second row presents results from SDE-SR, FASR (ours), and GT images.

ACKNOWLEDGMENTS

This study was financed in part by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES)* - Finance Code 001, and in part by the *Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)* (# 315409/2023-1). We thank the support of NVIDIA Corporation with the donation of the Quadro RTX 8000 GPU used for this research.

REFERENCES

[1] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1167–1183, 2002.

[2] J. Jiang, C. Wang, X. Liu, and J. Ma, "Deep learning-based face super-resolution: A survey," *ACM Computing Surveys (CSUR)*, vol. 55, no. 1, pp. 1–36, 2021.

[3] S. Zhu, S. Liu, C. C. Loy, and X. Tang, "Deep cascaded bi-network for face hallucination," in *European Conference on Computer Vision (ECCV)*, pp. 614–630, 2016.

[4] V. Nascimento *et al.*, "Combining attention module and pixel shuffle for license plate super-resolution," in *Conference on Graphics, Patterns and Images (SIBGRAPI)*, pp. 228–233, Oct 2022.

[5] X. Yu, B. Fernando, R. Hartley, and F. Porikli, "Super-resolving very low-resolution face images with supplementary attributes," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 908–917, 2018.

[6] M. Santos *et al.*, "Defying limits: Super-resolution refinement with diffusion guidance," in *International Conference on Computer Vision Theory and Applications (VISAPP)*, pp. 426–434, 2024.

[7] C. Saharia *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," in *International Conf. on Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 36 479–36 494, 2022.

[8] J. Richter, S. Frintrop, and T. Gerkmann, "Audio-visual speech enhancement with score-based generative models," in *ITG Conference on Speech Communication*, pp. 275–279, 2023.

[9] Z. Huang, K. C. Chan, Y. Jiang, and Z. Liu, "Collaborative diffusion for multi-modal face generation and editing," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6080–6090, 2023.

[10] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning*, pp. 2256–2265, 2015.

[11] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *International Conference on Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 6840–6851, 2020.

[12] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 1–13, 2019.

[13] ——, "Improved techniques for training score-based generative models," *Advances in neural information processing systems*, vol. 33, pp. 12 438–12 448, 2020.

[14] Y. Song *et al.*, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations (ICLR)*, pp. 1–36, May 2021.

[15] C. Niu *et al.*, "Permutation invariant graph generation via score-based generative modeling," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 108, pp. 4474–4484, Aug 2020.

[16] R. Cai, G. Yang, H. Averbuch-Elor, Z. Hao, S. Belongie, N. Snavely, and B. Hariharan, "Learning gradient fields for shape generation," in *European Conference on Computer Vision (ECCV)*, pp. 364–381, 2020.

[17] Y. Song, L. Shen, L. Xing, and S. Ermon, "Solving inverse problems in medical imaging with score-based generative models," in *International Conference on Learning Representations (ICLR)*, pp. 1–18, 2022.

[18] Z. Zhang *et al.*, "Sine: Single image editing with text-to-image diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6027–6037, 2023.

[19] C. Saharia *et al.*, "Image super-resolution via iterative refinement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4713–4726, 2023.

[20] H. Li, Y. Yang, M. Chang, S. Chen, H. Feng, Z. Xu, Q. Li, and Y. Chen, "SRDiff: Single image super-resolution with diffusion probabilistic models," *Neurocomputing*, vol. 479, pp. 47–59, 2022.

[21] M. Santos *et al.*, "Face super-resolution using stochastic differential equations," in *Conference on Graphics, Patterns and Images (SIBGRAPI)*, pp. 216–221, Oct 2022.

[22] M. Suin *et al.*, "Diffuse and restore: A region-adaptive diffusion model for identity-preserving blind face restoration," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6343–6352, 2024.

[23] A. Jolicoeur-Martineau, K. Li, R. Piché-Taillefer, T. Kachman, and I. Mitliagkas, "Gotta go fast when generating data with score-based models," *arXiv preprint arXiv:2105.14080*, 2021.

[24] A. Vahdat, K. Kreis, and J. Kautz, "Score-based generative modeling in latent space," in *International Conference on Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 11 287–11 302, 2021.

[25] C. Meng *et al.*, "On distillation of guided diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14 297–14 306, 2023.

[26] B. D. Anderson, "Reverse-time diffusion equation models," *Stochastic Processes and their Applications*, vol. 12, no. 3, pp. 313–326, 1982.

[27] P. Kloeden and E. Platen, *The Numerical Solution of Stochastic Differential Equations*, vol. 23. Springer, Jan 2011.

[28] S. Särkkä and A. Solin, *Applied stochastic differential equations*, vol. 10. Cambridge University Press, 2019.

[29] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural computation*, vol. 23, no. 7, pp. 1661–1674, 2011.

[30] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International Conference on Machine Learning (ICML)*, pp. 8162–8171, 2021.

[31] AssemblyAI, "How Imagen Actually Works," 2023. [Online]. Available: https://www.assemblyai.com/blog/how-imagen-actually-works/

[32] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4396–4405, 2019.

[33] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *International Conference on Computer Vision*, 2015.

[34] J. Neves, J. Moreno, and H. Proença, "QUIS-CAMPI: an annotated multi-biometrics data feed from surveillance scenarios," *IET Biometrics*, vol. 7, no. 4, pp. 371–379, 2018.

[35] M. Kim, A. K. Jain, and X. Liu, "AdaFace: Quality adaptive margin for face recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[37] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[38] S. Gao, X. Liu, B. Zeng, S. Xu, Y. Li, X. Luo, J. Liu, X. Zhen, and B. Zhang, "Implicit diffusion models for continuous super-resolution," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 021–10 030, 2023.