Defying Limits: Super-Resolution Refinement with Diffusion Guidance

Marcelo dos Santos[1][a], João C. R. Neves [2][b], Hugo Proença[2][c] and David Menotti[1][d]
[1]*Department of Informatics, Federal University of Paraná, Curitiba, Brazil*
[2]*Instituto de Telecomunicações, University of Beira Interior, Covilhã, Portugal*
{*msantos, menotti*}*@inf.ufpr.br,* {*jcneves, hugomcp*}*@di.ubi.pt*

Keywords: Super-Resolution, Face Recognition, Diffusion Models, Diffusion Guidance.

Abstract: Due to the growing number of surveillance cameras and rapid technological advancement, facial recognition algorithms have been widely applied. However, their performance decreases in challenging environments, such as those involving surveillance cameras with low-resolution images. To address this problem, in this paper, we introduce SRDG, a super-resolution approach supported by two state-of-the-art methods: diffusion models and classifier guidance. The diffusion process reconstructs the image, and the classifier refines the image reconstruction based on a set of facial attributes. This combination of models is capable of working with images with a very limited resolution (8×8 and 16×16), being suitable for surveillance scenarios where subjects are typically distant from the camera. The experimental validation of the proposed approach shows that super-resolution images exhibit enhanced details and improved visual quality. More importantly, when using our super-resolution algorithm, the facial discriminability of images is improved compared to state-of-the-art super-resolution approaches, resulting in a significant increase in face recognition accuracy. To the best of our knowledge, this is the first time classifier guidance has been applied to refine super-resolution results of images from surveillance cameras. Source code is available at https://github.com/marcelowds/SRDG.

1 INTRODUCTION

Super-resolution (SR) refers to the process of transforming a low-resolution (LR) degraded image into a higher resolution and less noisy image, aiming to enhance the visual information contained in the LR image [Abiantun et al., 2019].

For surveillance environments and real-world scenarios, the performance of super-resolution (SR) and face recognition (FR) algorithms, such as AdaFace [Kim et al., 2022b] and ArcFace [Deng et al., 2019], falls drastically. The numerous challenges posed by factors such as pose, variations in lighting conditions, occlusions and other pertinent issues are the main contributors to this decline [Zhu et al., 2016].

The use of soft biometrics, such as gender, facial marks, age, and other characteristics, has the potential to improve facial recognition and super-resolution results [Lee et al., 2018, Li et al., 2020, Yu et al., 2018, Yu et al., 2020, Lu et al., 2018]. Considering that soft biometrics are available in many cases, this additional information is used in this work to augment the performance of SR algorithms. More specifically, we will use soft biometrics to simultaneously improve the quality of super-resolved images and the accuracy of face recognition methods.

Recently, numerous works that use diffusion models have emerged (as detailed in the surveys [Yang et al., 2022, Cao et al., 2022, Croitoru et al., 2022, Li et al., 2023]). These models employ the concept of perturbing data with different noise scales and train a neural network to predict the noise of the data. Once the neural network is trained, it becomes possible to perform reverse diffusion, removing noise and generating a specific data type.

An additional tool usually employed in diffusion models is classifier guidance, which utilizes the gradient of an attribute classifier combined with the score function (i.e., the gradient of the log probability density with respect to data) of a diffusion model to orient the reverse diffusion process [Nichol and Dhariwal, 2021]. This guidance allows the output to be directed to a pre-defined class [Song et al., 2021].

Based on these ideas, this paper addresses the challenges that SR and facial recognition algorithms face in surveillance environments. By combining the

[a] https://orcid.org/0000-0003-0960-2641
[b] https://orcid.org/0000-0003-0139-2213
[c] https://orcid.org/0000-0003-2551-8570
[d] https://orcid.org/0000-0003-2430-2030

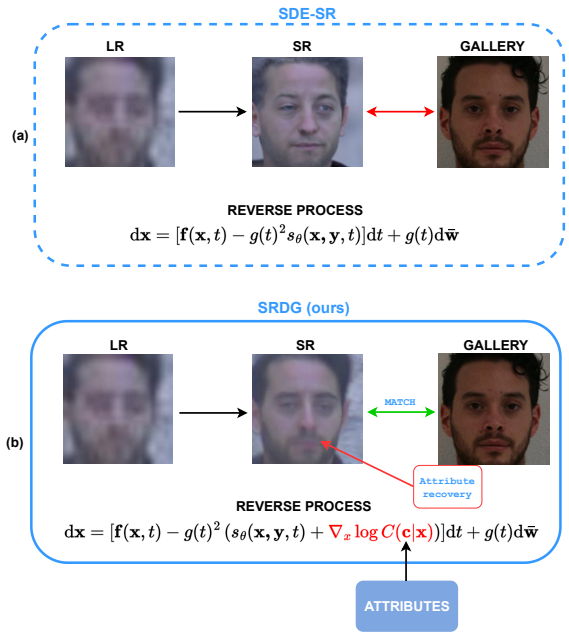

Figure 1: (a) Illustration of the conventional super-resolution algorithm based on stochastic differential equations SDE-SR [Santos et al., 2022]. (b) Our method: The classifier guidance approach is used to include complementary attributes for generating more detailed super-resolution images. With higher-quality images, face recognition can be performed more accurately.

data generation capabilities of diffusion models [Santos et al., 2022, Ho et al., 2020] with classifier guidance [Dhariwal and Nichol, 2021] (see Figure 1), we seek to enhance the quality of extremely LR images (8×8 and 16×16) obtained from surveillance cameras in unconstrained scenarios.

The main contribution of our work lies in employing soft biometrics as a source of information for the attribute classifier to guide the reverse diffusion process. The effectiveness of the method is assessed in the Quis-Campi dataset [Neves et al., 2018], which comprises realistic data from surveillance scenarios. The proposed approach yielded superior qualitative and quantitative results, as demonstrated by the visual quality of the images and by the metrics: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM). Additionally, our methodology excelled in face recognition metrics, such as Area Under the Curve (AUC) (1:1 verification protocol) and accuracy (1:N identification protocol).

This paper is structured in the following manner: Section 2 includes the related work, Section 3 introduces the proposed method, and Section 4 outlines our experiments and the corresponding results. The conclusions of the paper are outlined in Section 5.

2 RELATED WORK

One important precursor work in diffusion models was [Sohl-Dickstein et al., 2015], where considerations from non-equilibrium thermodynamics were used to generate images. From then on, two other models of importance were the Denoising Diffusion Probabilistic Models (DDPMs) [Ho et al., 2020] and Score-Based Generative Models (SGMs). In [Song et al., 2021], DDPM and SGD are generalized for continuous time steps and noise levels employing a Stochastic Differential Equation (SDE), giving rise to the models VP (Variation Preserving) and VE (Variation Exploding), respectively.

Diffusion models can be applied to data generation across diverse domains such as generation of audio [Chen et al., 2020], graphs [Niu et al., 2020] and shapes [Cai et al., 2020] as well as for image synthesis [Ho et al., 2020, Song and Ermon, 2019, Song et al., 2021]. For the image synthesis task, diffusion models provide more satisfactory image quality and training stability compared to Generative Adversarial Networks (GANs) [Dhariwal and Nichol, 2021]. Among other applications, domain translation can also be combined with diffusion models for text-to-image translation [Saharia et al., 2022].

Inspired by the DDPM diffusion model, SR3 [Saharia et al., 2021] transforms images with pure noise in SR images by conditioning a neural network on an LR input through a Markov chain. [Li et al., 2022] proposed SRDiff, which utilizes the same idea of SR3, but the difference is that the residual SR image is estimated, and the final SR image is obtained by adding the predicted SR residue to the original image upscaled. [Gao et al., 2023] is an improvement of SR3 and can perform SR with a continuous scale factor. The work [Santos et al., 2022] develops SDE-SR, which also performs SR using diffusion models but employing a SDE.

Despite the several advantages of diffusion models, such as data quality and training stability (unlike GANs), a weakness of these models is their high execution time. The works [Song et al., 2020, Jolicoeur-Martineau et al., 2021, Vahdat et al., 2021] have been dedicated to increase the efficiency of diffusion models while improving the quality of the resulting samples. [Meng et al., 2023] performs diffusion in specific tasks using as few as $2-4$ denoising steps.

Diffusion models can also be used as conditional generators. [Dhariwal and Nichol, 2021] describes a method for using gradients from a classifier to guide a diffusion model during sampling. This conditional generator method will be used in this work.

3 PROPOSED METHOD

Despite the low quality of data acquired in surveillance scenarios, specific attributes, such as gender, the use of eyeglasses, beard, and others, can sometimes be determined, see Figure 2. In this manner, these additional pieces of information can be utilized to perform SR and facial recognition. Next we show how the stochastic differential equations-based SR technique can be further improved to perform SR by incorporating complementary attributes. We will refer to our method as SRDG (Super-Resolution with Diffusion Guidance).

Figure 2: An image captured by a surveillance camera enables an expert to gather key attributes such as gender, the presence of a beard, eyeglasses, and other characteristics during a forensic analysis.

In [Song et al., 2021], diffusion models are modeled as a continuous diffusion process $\{x(t)\}_{t=0}^{T}$ by the Itô SDE

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

where $\mathbf{f}(\mathbf{x},t)$ is the drift coefficient, $g(t)$ is a diffusion coefficient, and $\mathbf{w}$ is a Wiener process. For more details about Itô SDE and Wiener process, see [Kloeden and Platen, 2011, Särkkä and Solin, 2019]. In [Anderson, 1982], it was shown that it is possible to reverse the diffusion process (Eq. 1) using another diffusion process given by

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

where $\mathrm{d}\bar{\mathbf{w}}$ is a Wiener process running backwards in time.

Similar to [Santos et al., 2022], here we consider $\mathbf{x}$ as the images to be denoised and $\mathbf{y}$ as the LR images. A neural network $s_\theta(\mathbf{x},\mathbf{y},t)$ conditioned on $\mathbf{x},\mathbf{y},t$ is used to approximate $\nabla_{\mathbf{x}}\log p_t(\mathbf{x})$. This is performed by optimizing the loss function [Vincent, 2011]

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

In this work, we consider $p_t(\mathbf{x}|\mathbf{c})$ on the reverse process as dependent on $\mathbf{x}$ and conditioned to the class $\mathbf{c}$ to which the image belongs. In this case, using Bayes' rule, we have

$$\nabla_{\mathbf{x}}\log p_t(\mathbf{x}|\mathbf{c}) = \nabla_{\mathbf{x}}\log p_t(\mathbf{x}) + \nabla_{\mathbf{x}}\log p_t(\mathbf{c}|\mathbf{x}). \tag{4}$$

But $\nabla_{\mathbf{x}}\log p_t(\mathbf{x})$ is already approximated by $s_\theta(\mathbf{x},\mathbf{y},t)$ and $p_t(\mathbf{c}|\mathbf{x})$ is a time dependent classifier C. Therefore, the reverse process given by Equation 2 becomes

$$\mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x},t) - g(t)^2\left(s_\theta(\mathbf{x},\mathbf{y},t) + h\nabla_{\mathbf{x}}\log \mathrm{C}(\mathbf{c}|\mathbf{x})\right)\right]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}. \tag{5}$$

Hence, with a classifier $\mathrm{C}(\mathbf{c}|\mathbf{x})$ trained on noisy images, it is possible to condition the image generation of the reverse process. Details about the dataset utilized to train the classifier, its architecture and training parameters are given in Subsections 4.1 and 4.2. Note that, as the reverse process is already conditioned by the LR image $\mathbf{y}$, and as we are interested in a refinement of the SR images, the class $\mathbf{c}$ and image $\mathbf{y}$ must be coherent, since classifier-guided diffusion sampling can be interpreted as attempting to confuse an image classifier with a gradient-based adversarial attack [Ho and Salimans, 2022]. So if LR face image $\mathbf{y}$ has eyeglasses, class $\mathbf{c}$ must also have the glasses attribute defined as True. Similar to other works, the classifier's gradient will be scaled by a constant factor $h > 1$, which is responsible for generating high-quality and less diverse images [Kim et al., 2022a, Ho and Salimans, 2022, Dhariwal and Nichol, 2021].

We are also following the VE (Variation Exploding) configuration defined in [Song et al., 2021] and [Santos et al., 2022]. In this case, $\mathbf{f}(\mathbf{x},t)$ and $g(t)$ are given respectively by

$$\mathbf{f}(\mathbf{x},t) = \mathbf{0}, \quad g(t) = \sqrt{\frac{\mathrm{d}\sigma^2(t)}{\mathrm{d}t}}. \tag{6}$$

Here we will use the same $\sigma(t)$ defined in [Song and Ermon, 2019] and given by $\sigma(t) = \sigma_{min}(\sigma_{max}/\sigma_{min})^t$. To perform the training we must have $p(\mathbf{x}(t)|\mathbf{x}(0))$ to compute the loss function (Equation 3). The mean and covariance of $p(\mathbf{x}_t|\mathbf{x}_0)$ are given by [Kloeden and Platen, 2011, Song et al., 2021]

$$\boldsymbol{\mu}(t) = \mathbf{x}(0),\ \boldsymbol{\Sigma}(t) = [\sigma^2(t) - \sigma^2(0)]\mathbf{I}, \tag{7}$$

so the term $\nabla_{\mathbf{x}}\log p(\mathbf{x}(t)|\mathbf{x}(0))$ can be analytically computed in Equation 3. Once we have trained the neural network $s_\theta(\mathbf{x},\mathbf{y},t)$ and the classifier $C(\mathbf{c}|\mathbf{x})$, it is possible to obtain SR images by performing the reverse diffusion process. In other words, starting with a pure noisy image $\mathbf{x}_T$ at $t = T$, we solve Equation 5 using Euler's method, and we obtain, at $t = 0$, the SR image $\mathbf{x}_0$ in a predefined class $\mathbf{c}$.

4 EXPERIMENTS AND RESULTS

4.1 Datasets

In this study, three distinct datasets were employed: (i) FFHQ [Karras et al., 2019] for training the SR model, (ii) CelebA [Liu et al., 2015] for classifier training, and (iii) Quis-Campi [Neves et al., 2018] for method validation and fine-tuning of the SR base model (SR model without classifier guidance).

Regarding the Quis-Campi dataset, 90 identities were considered for the method validation. For each identity, we used a mugshot frontal acquired in a controlled environment as a gallery image and five probe images from a surveillance camera. The remaining probe images where a face was visible were used to fine-tune the SR method.

4.2 Architectures and Training

Similar to other diffusion models, the network architecture of the main SR model is based on the U-net architecture [Ho et al., 2020] but adapted to receive the LR image $\mathbf{y}$, concatenated with the image to be denoised $\mathbf{x}_t$. Following [Song et al., 2021] and [Santos et al., 2022], we set the parameters of $\sigma(t)$ equals to $\sigma_{min} = 0.01$ and $\sigma_{max} = 348$. For the model training, Adam optimizer was used with a warm-up of 5000 steps and a learning rate of 2×10^{-4}.

The training process of the SR base model (i.e., the SR model without the classifier) included two key stages. Initially, high-resolution (HR) images from the FFHQ dataset were utilized. To mimic LR scenarios, the images were downscaled by factors of 8× and 16×, generating pairs of LR and HR images. The algorithm was then trained across 10^6 steps using these paired images. Subsequently, the SR base model underwent fine-tuning through an additional 10^5 training steps using images from the Quis-Campi dataset. For SR image generation, the total number of time steps was set in 2000.

During the reconstruction phase, the diffusion guidance is performed with the Densenet classifier [Huang et al., 2017], adapted to incorporate the time variable, which correlates with the noise level present in the image. The training took place for 50 epochs, with a learning rate of 10^{-3}, a batch size of 4, and utilizing AdamW optimizer. The scaling factor for the classifier gradient was configured to be $h = 50$.

The accuracy of the classifier is dependent on the diffusion time. Figure 3 shows the accuracy of the classifier as a function of time for three attributes: gender, beard and eyeglasses. As can be seen, the classifier achieves accuracy higher than 88% for shorter time intervals. However, as the time increases, the accuracy rapidly declines due to the predominant noise in the image.

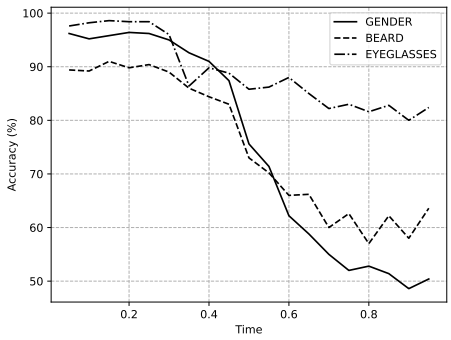

Figure 3: Classifier accuracy as a function of time for the attributes gender, beard and eyeglasses.

4.3 Feature Extraction

To construct a feature vector, a 512-dimensional descriptor was extracted from images using the ResNet backbone [He et al., 2016] with the modifications performed by [Kim et al., 2022b] and pre-trained on CASIA-WebFace [Yi et al., 2014]. For the face recognition task, we relied on AdaFace [Kim et al., 2022b], and image descriptors were compared using the cosine similarity metric.

4.4 Experiments

Before the generation of SR images, the attributes are determined through forensic analysis, although they can also be obtained using the classifier trained with LR images.

In order to assess the significance of soft attributes in both SR and facial recognition, we evaluated our approach on 8×8 and 16×16 images and used an upsampling factor of 16× and 8× to obtain 128×128 super-resolved images, respectively.

In the recognition task, the super-resolved images are matched against the gallery images, whereas within the scope of the SR task, the recovered images are compared with the original probe images. Our method is compared against the methods SR3 [Saharia et al., 2021], IDM [Gao et al., 2023], and the baseline SDE-SR [Santos et al., 2022].

4.5 Results

Given the scarcity of SR algorithms that perform 16× upsampling, our comparison for this scale is solely

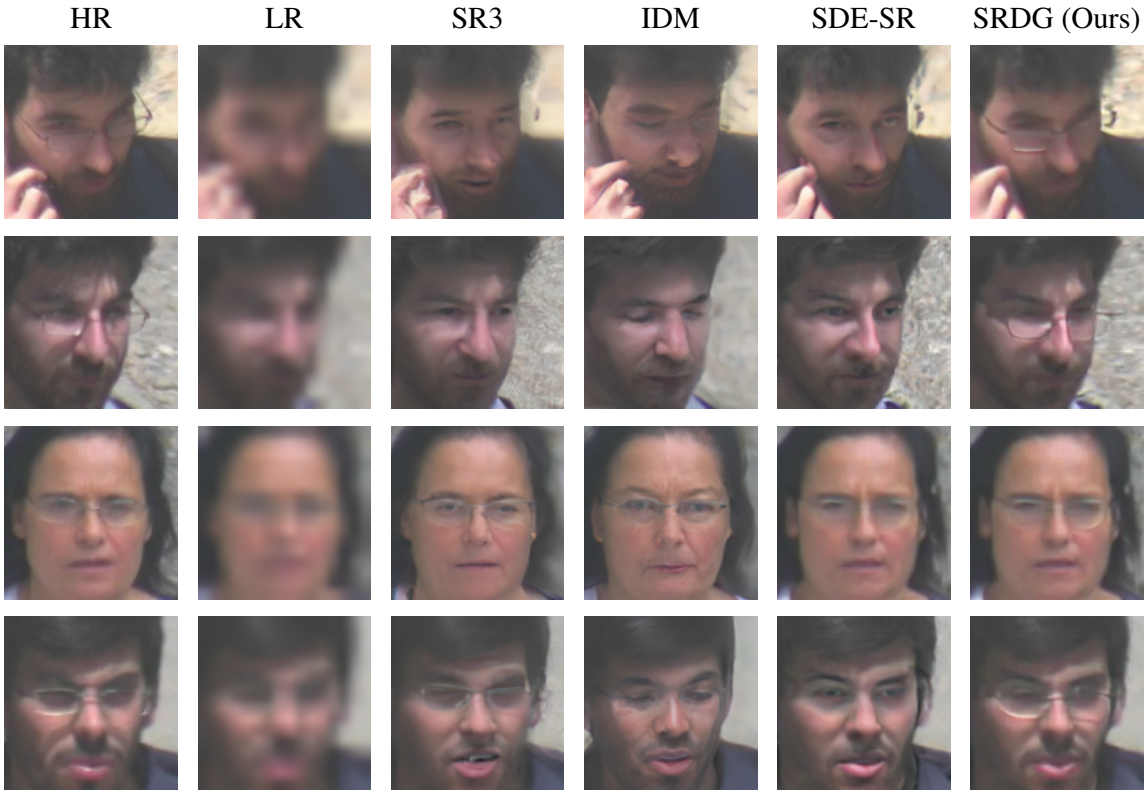

Figure 4: 8× super-resolution results with the use of soft biometrics.

conducted with the SR3 and SDE-SR algorithms, which were retrained for 16× upsampling. The quantitative results of the SR process are presented in Table 1, highlighting the superior performance of our algorithm across PSNR and SSIM metrics. In addition, Table 2 reports the performance of a state-of-the-art face recognition algorithm when provided with original LR and SR images. The results show a significant difference in face recognition performance when SR techniques are used in surveillance scenarios, which justifies using these algorithms. Regarding the comparison between SR strategies, our approach surpasses the remaining methods, evidencing the advantages of the classifier guidance process.

Figures 4 and 5 show qualitative results of 8× and 16× SR algorithms, respectivelly. As can be seen, our approach can recover even the finest details as eyeglasses contours and retain the discriminant visual features of the face, explaining the quantitative improvements across all face recognition metrics.

4.6 Ablation Study

Here, our method is compared with the SDE-SR baseline. Tests on our algorithm were conducted using

Table 1: PSNR and SSIM results for the Quis-Campi dataset with upscaling factor of 8× and 16×.

	PSNR ↑		SSIM ↑	
SR Method	8×	16×	8×	16×
SR3	30.71	23.57	0.86	0.65
IDM	26.25	-	0.78	-
SDE-SR	30.34	24.26	0.84	0.69
SRDG (Ours)	32.46	27.49	0.88	0.81

only one attribute (gender), and also three attributes (gender, beard and eyeglasses). For each of these cases, the models with and without fine-tuning were tested.

Table 3 shows the ablation study performed to validate the use of soft biometrics on SRDG (upsampling factors of 8× and 16×), applied to face recognition tasks. As can be seen, higher values for the AUC metric are obtained with our method, i.e., when attributes are used (one or three), and this holds for both the case with fine-tuning and the case without fine-tuning.

Concerning recognition accuracy, the highest values without fine-tuning are obtained with our method. However, upon fine-tuning, the impact of attributes is more relevant only for an upsampling factor of 8×. For more reliable results from SDE-SR and SRDG,

Figure 5: 16× super-resolution results with the use of soft biometrics.

methods to minimize the distortion of the person's identity are necessary during the image reconstruction since SR algorithms are ill-posed problems [Baker and Kanade, 2002].

5 CONCLUSIONS AND FUTURE WORK

Conventional SR techniques based on SDEs depend exclusively on the score function for creating a SR image through reverse diffusion. In contrast, this work introduces a SR method that relies on complementary attributes to enhance the quality of super-resolved images. Our approach employs the gradient of an attribute classifier to guide the reverse process. During the reconstruction process, our method can not only recover discernible features such as facial traits but also subtle characteristics that improve the discriminability for face recognition. A significant advantage of using classifier guidance is that the SR model does not need to be retrained, which provides practicality to the method.

The efficacy of our approach in restoring uncommonly recovered structures and local features by SR algorithms has been demonstrated by the evaluation of the proposed approach with respect to image quality and face recognition metrics.

Regarding image quality, our approach is capable of recovering finer details from extremely low-resolution images (8×8 or 16×16), which has been confirmed by improvement in quantitative (PSNR and SSIM) and qualitative (visually) results over competing SR approaches.

The experiments on the Quis-Campi dataset evidence a significant improvement in the recognition performance when using the super-resolved images produced by our approach, indicating that our algorithm has the potential for working in surveillance scenarios where the data resolution is typically very small.

Given that the initial attributes are extracted from LR images, uncertainties in their predictions may propagate to SR images, resulting in inaccurate outcomes. Therefore, employing SRDG in scenarios with low-accuracy attribute predictions should be avoided. Furthermore, SR images suffer from bias issues due to the ill-posedness of SR methods. Consequently, although our method provides superior re-

Table 2: The 1:1 verification and 1:N identification (Rank-1, Rank-5 and Rank-10) results from 8× and 16× super-resolution for Quis-Campi dataset with AdaFace FR model. The superscript † stands for fine-tuned.

	AUC		Rank-1 (%)		Rank-5 (%)		Rank-10 (%)	
SR Method	8×	16×	8×	16×	8×	16×	8×	16×
LR	0.816	0.610	23.78	5.11	46.89	16.44	58.67	24.44
SR3	0.914	0.702	45.78	7.78	69.56	23.11	79.77	34.44
IDM	0.885	-	28.22	-	56.44	-	70.00	-
SDE-SR	0.917	0.697	50.00	9.33	72.67	24.00	81.56	36.67
SRDG (Ours)	0.920	0.696	49.33	10.00	73.11	25.56	82.00	36.00
SDE-SR†	0.922	0.812	57.78	26.00	76.22	50.44	83.56	63.78
SRDG† (Ours)	0.929	0.818	57.11	24.67	79.11	48.44	85.56	64.44

Table 3: Ablation study for the 1:1 verification and 1:N identification (Rank-1, Rank-5 and Rank-10) results from 8× and 16× super-resolution for Quis-Campi dataset with AdaFace FR model. The superscript † stands for fine-tuned (FT).

			AUC		Rank-1 (%)		Rank-5 (%)		Rank-10 (%)	
FT	# Attrs	SR Method	8×	16×	8×	16×	8×	16×	8×	16×
✗	-	SDE-SR	0.917	0.697	50.00	9.33	72.67	24.00	81.56	36.67
✗	1	SRDG (Ours)	0.918	0.701	50.00	10.44	72.22	24.44	81.78	36.22
✗	3	SRDG (Ours)	0.920	0.696	49.33	10.00	73.11	25.56	82.00	36.00
✓	-	SDE-SR†	0.922	0.812	57.78	26.00	76.22	50.44	83.56	63.78
✓	1	SRDG† (Ours)	0.926	0.814	58.00	23.78	77.11	49.11	82.67	63.56
✓	3	SRDG† (Ours)	0.929	0.818	57.11	24.67	79.11	48.44	85.56	64.44

sults for face recognition, it can only be applied in real situations after these bias problems are resolved. In future works, a method must be developed to minimize distortions in the person's identity when working with diffusion models.

ACKNOWLEDGEMENTS

This work was partly supported by the Coordination for the Improvement of Higher Education Personnel (CAPES) (*Programa de Cooperação Acadêmica em Segurança Pública e Ciências Forenses # 88887.619562/2021-00*), partly by the National Council for Scientific and Technological Development (CNPq) (# 308879/2020-1), and partly by NOVA LINCS (UIDP/04526/2020) with the financial support of the Foundation for Science and Technology (FCT).

REFERENCES

Abiantun, R., Juefei-Xu, F., Prabhu, U., and Savvides, M. (2019). SSR2: Sparse signal recovery for single-image super-resolution on faces with extreme low resolutions. *Pattern Recognition*, 90:308–324.

Anderson, B. D. (1982). Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326.

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.

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.

Cao, H., Tan, C., Gao, Z., Chen, G., Heng, P.-A., and Li, S. Z. (2022). A survey on generative diffusion model. *arXiv preprint arXiv:2209.02646*.

Chen, N., Zhang, Y., Zen, H., Weiss, R. J., Norouzi, M., and Chan, W. (2020). Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*.

Croitoru, F.-A., Hondru, V., Ionescu, R. T., and Shah, M. (2022). Diffusion models in vision: A survey. *arXiv preprint arXiv:2209.04747*.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 4690–4699.

Dhariwal, P. and Nichol, A. (2021). Diffusion models beat GANs on image synthesis. In *International Conference on Neural Information Processing Systems (NeurIPS)*, volume 34, pages 8780–8794.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10021–10030.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of*

the IEEE conference on Computer Vision and Pattern Recognition, pages 770–778.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In *International Conference on Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851.

Ho, J. and Salimans, T. (2022). Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4700–4708.

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 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.

Kim, H., Kim, S., and Yoon, S. (2022a). Guided-tts: A diffusion model for text-to-speech via classifier guidance. In *International Conference on Machine Learning*, pages 11119–11133. PMLR.

Kim, M., Jain, A. K., and Liu, X. (2022b). Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Kloeden, P. and Platen, E. (2011). *The Numerical Solution of Stochastic Differential Equations*, volume 23. Springer.

Lee, C.-H., Zhang, K., Lee, H.-C., Cheng, C.-W., and Hsu, W. (2018). Attribute augmented convolutional neural network for face hallucination. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition workshops*, pages 721–729.

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.

Li, M., Zhang, Z., Yu, J., and Chen, C. W. (2020). Learning face image super-resolution through facial semantic attribute transformation and self-attentive structure enhancement. *IEEE Transactions on Multimedia*, 23:468–483.

Li, X., Ren, Y., Jin, X., Lan, C., Wang, X., Zeng, W., Wang, X., and Chen, Z. (2023). Diffusion models for image restoration and enhancement–a comprehensive survey. *arXiv preprint arXiv:2308.09388*.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Lu, Y., Tai, Y.-W., and Tang, C.-K. (2018). Attribute-guided face generation using conditional cyclegan. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 282–297.

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

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.

Nichol, A. and Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. *arXiv preprint arXiv:2102.09672*.

Niu, C. et al. (2020). Permutation invariant graph generation via score-based generative modeling. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108, pages 4474–4484.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494.

Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., and Norouzi, M. (2021). Image super-resolution via iterative refinement. *arXiv preprint*, arXiv:2104.07636:1–28. Google Research.

Santos, M. D., Laroca, R., Ribeiro, R. O., Neves, J., Proença, H., and Menotti, D. (2022). Face super-resolution using stochastic differential equations. In *2022 35th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, volume 1, pages 216–221. IEEE.

Särkkä, S. and Solin, A. (2019). *Applied stochastic differential equations*, volume 10. Cambridge University Press.

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

Song, J., Meng, C., and Ermon, S. (2020). Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y. and Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. In *International Conference on Neural Information Processing Systems (NeurIPS)*, pages 1–13.

Song, Y. et al. (2021). Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, pages 1–36.

Vahdat, A., Kreis, K., and Kautz, J. (2021). Score-based generative modeling in latent space. In *International Conference on Neural Information Processing Systems (NeurIPS)*, volume 34, pages 11287–11302.

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

Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Shao, Y., Zhang, W., Cui, B., and Yang, M.-H. (2022). Diffusion models: A comprehensive survey of methods and applications. *arXiv preprint arXiv:2209.00796*.

Yi, D., Lei, Z., Liao, S., and Li, S. Z. (2014). Learn-

ing face representation from scratch. *arXiv preprint arXiv:1411.7923*.

Yu, X., Fernando, B., Hartley, R., and Porikli, F. (2018). Super-resolving very low-resolution face images with supplementary attributes. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 908–917.

Yu, X., Fernando, B., Hartley, R., and Porikli, F. (2020). Semantic face hallucination: Super-resolving very low-resolution face images with supplementary attributes. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 42(11):2926–2943.

Zhu, S., Liu, S., Loy, C. C., and Tang, X. (2016). Deep cascaded bi-network for face hallucination. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 614–630. Springer.