Face Super-Resolution Using Stochastic Differential Equations

Marcelo dos Santos[1], Rayson Laroca[1], Rafael O. Ribeiro[2], João Neves[3], Hugo Proença[3], David Menotti[1]

Department of Informatics, Federal University of Paraná, Curitiba, Brazil[1]
National Institute of Criminalistics, Brazilian Federal Police, Brasília, Brazil[2]
Instituto de Telecomunicações, University of Beira Interior, Covilhã, Portugal[3]

SIBGRAPI - Conference on Graphics, Patterns and Images
October 24-27, 2022

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Outline

1. Introduction
2. Theoretical Background
3. Experiments and Results
4. Conclusions

Introduction

Diffusion Models

Forward Diffusion (data $\rightarrow$ noise)

Reverse Diffusion (noise $\rightarrow$ data)

Introduction

Score-based generative models have been successfully applied to

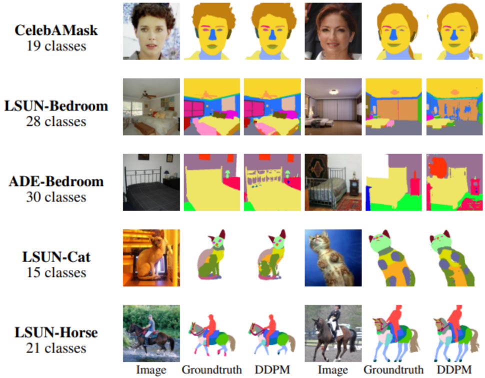
- Image Synthesis
- Semantic Segmentation
- 3D Shape Generation
- Text-to-Image
- Video Synthesis
- Medical Imaging

Song et al., Score-based generative modeling through stochastic differential equations, ICLR, 2021

Introduction

Score-based generative models have been successfully applied to

- Image Synthesis
- Semantic Segmentation
- 3D Shape Generation
- Text-to-Image
- Video Synthesis
- Medical Imaging



Baranchuk et al., Label-Efficient Semantic Segmentation with Diffusion Models, ICLR, 2022

Introduction

Score-based generative models have been successfully applied to

- Image Synthesis
- Semantic Segmentation
- 3D Shape Generation
- Text-to-Image
- Video Synthesis
- Medical Imaging

Zhou et al., 3D Shape Generation and Completion through Point-Voxel Diffusion, ICCV, 2021

Introduction

Score-based generative models have been successfully applied to

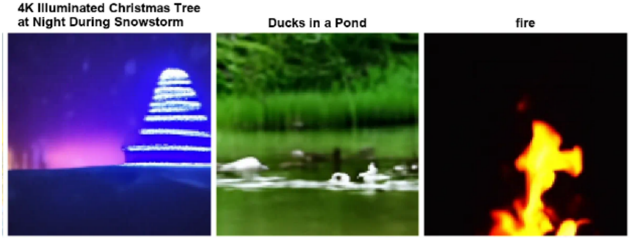
- Image Synthesis
- Semantic Segmentation
- 3D Shape Generation
- Text-to-Image
- Video Synthesis
- Medical Imaging

Saharia et al., **Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding**, arXiv, 2022

Introduction

Score-based generative models have been successfully applied to

- Image Synthesis
- Semantic Segmentation
- 3D Shape Generation
- Text-to-Image
- Video Synthesis
- Medical Imaging

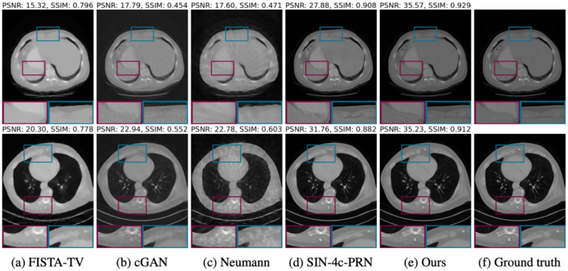


https://video-diffusion.github.io/

Introduction

Score-based generative models have been successfully applied to

- Image Synthesis
- Semantic Segmentation
- 3D Shape Generation
- Text-to-Image
- Video Synthesis
- Medical Imaging



Song et al., Solving Inverse Problems in Medical Imaging with Score-Based Generative Models, ICLR, 2022

Theoretical Background

Stochastic Differential Equations

Forward SDE (data $\rightarrow$ noise)

$$\mathbf{x}(0) \longrightarrow \mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w} \longrightarrow \mathbf{x}(T)$$

score function

$$\mathbf{x}(0) \longleftarrow \mathrm{d}\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g(t)^2 \boxed{\nabla_{\mathbf{x}} \log p_t(\mathbf{x})}]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}} \longleftarrow \mathbf{x}(T)$$

Reverse SDE (noise $\rightarrow$ data)

Stochastic Differential Equations

Equivalence of SMLD and DDPM models with SDEs

Diffusion models

- Score Matching with Langevin Dynamics (SMLD)
- Denoising Diffusion Probabilistic Models (DDPM)

Stochastic Differential Equations

Equivalence of SMLD and DDPM models with SDEs

Diffusion models

- Score Matching with Langevin Dynamics (SMLD)
- Denoising Diffusion Probabilistic Models (DDPM)

Song et al. [1] generalizes diffusion models to SDEs

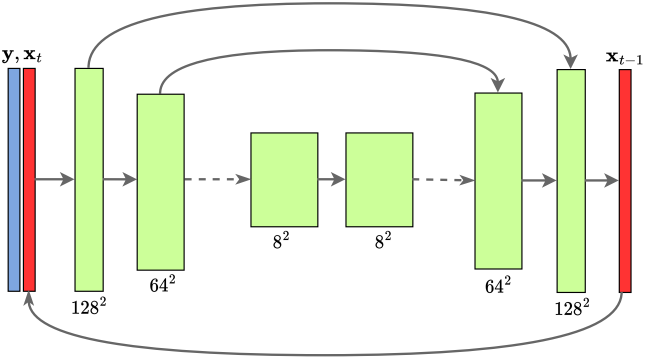
- SDE-VE (SMLD)
- SDE-VEcs (correction step)
- SDE-VP (DDPM)
- SDE-subVP

Super-Resolution Loss function

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

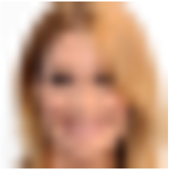
Super-Resolution Architecture

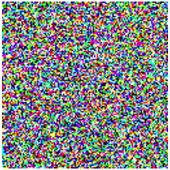
U-Net architecture with ResNet blocks and self-attention layers

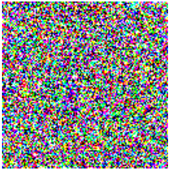


How it works

$\mathbf{y}$ $\qquad$ $\mathbf{x}_T$ $\qquad \Longrightarrow \qquad$ $\mathbf{x}_{T-1}$

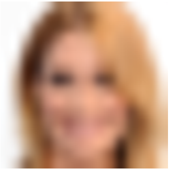




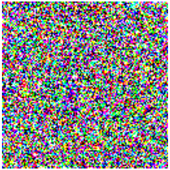


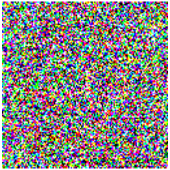
How it works

$\mathbf{y}$



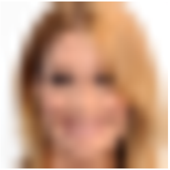
$\mathbf{x}_t \Longrightarrow \mathbf{x}_{t-1}$

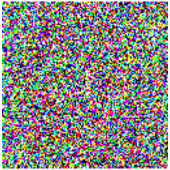


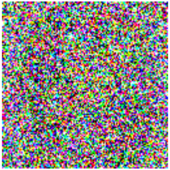


How it works

$\mathbf{y}$ $\qquad$ $\mathbf{x}_t$ $\qquad \Longrightarrow \qquad$ $\mathbf{x}_{t-1}$

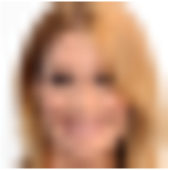




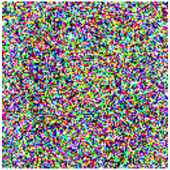


How it works

$\mathbf{y}$

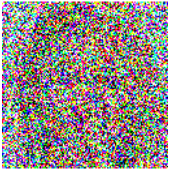


$\mathbf{x}_t$



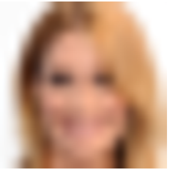
$\Longrightarrow$

$\mathbf{x}_{t-1}$

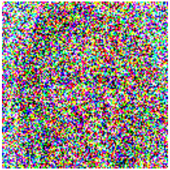


How it works

$\mathbf{y}$

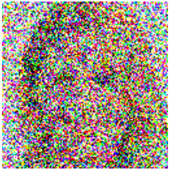


$\mathbf{x}_t$



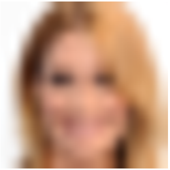
$\Longrightarrow$

$\mathbf{x}_{t-1}$



How it works

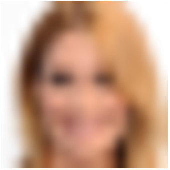
$\mathbf{y}$



$\mathbf{x}_t \Longrightarrow \mathbf{x}_{t-1}$

How it works

$\mathbf{y}$



$\mathbf{x}_t$ $\Longrightarrow$

$\mathbf{x}_{t-1}$

How it works

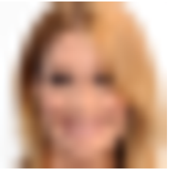
$\mathbf{y}$ $\mathbf{x}_t$ $\Longrightarrow$ $\mathbf{x}_{t-1}$

How it works

$\mathbf{y}$ $\quad$ $\mathbf{x}_t$ $\Longrightarrow$ $\mathbf{x}_{t-1}$

How it works

$\mathbf{y}$



$\mathbf{x}_1$

$\Longrightarrow$

$\mathbf{x}_0$

Algorithm

Algorithm 1 Predictor-Corrector (PC) sampling

N: Number of discretization steps for the reverse-time SDE
M: Number of correction steps

Initialize $\mathbf{x}_T \sim p_T(\mathbf{x})$
for $i = N - 1$ **to** 0 **do**
    $\mathbf{x}_i \leftarrow \text{Predictor}(\mathbf{x}_{i+1})$
    for $j = 1$ **to** M **do**
        $\mathbf{x}_i \leftarrow \text{Corrector}(\mathbf{x}_i)$ ▷ Internal parameter r related to image smoothness
    end for
end for
return $\mathbf{x}_0$

Image smoothness

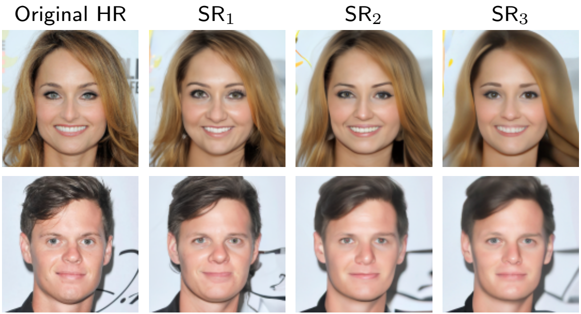


Figure: Samples obtained increasing r from left to right. For Samples SR_1, SR_2 and SR_3 the values of r are 0.10, 0.30 and 0.52 for the first row and 0.12, 0.32 and 0.39 for the second row. Higher values of r yield smoother images and larger values of PSNR (on average).

Experiments and Results

Experiments

- Training dataset: Flickr - Faces - HQ (FFHQ)

Experiments

- Training dataset: Flickr - Faces - HQ (FFHQ)
- LR images: obtained by downsampling HR images by a factor of 8

Experiments

- Training dataset: Flickr - Faces - HQ (FFHQ)
- LR images: obtained by downsampling HR images by a factor of 8
- Testing dataset: CelebA-HQ

Experiments

- Training dataset: Flickr - Faces - HQ (FFHQ)
- LR images: obtained by downsampling HR images by a factor of 8
- Testing dataset: CelebA-HQ
- Features extraction of HR and SR images with VGG-Face

Experiments

- Training dataset: Flickr - Faces - HQ (FFHQ)
- LR images: obtained by downsampling HR images by a factor of 8
- Testing dataset: CelebA-HQ
- Features extraction of HR and SR images with VGG-Face
- Metrics: SSIM, PSNR and Consistency
    - Cosine similarity CS between feature vectors

Experiments

- Training dataset: Flickr - Faces - HQ (FFHQ)
- LR images: obtained by downsampling HR images by a factor of 8
- Testing dataset: CelebA-HQ
- Features extraction of HR and SR images with VGG-Face
- Metrics: SSIM, PSNR and Consistency
  - Cosine similarity CS between feature vectors
- Influence of image smoothness on PSNR and CS metrics

Results

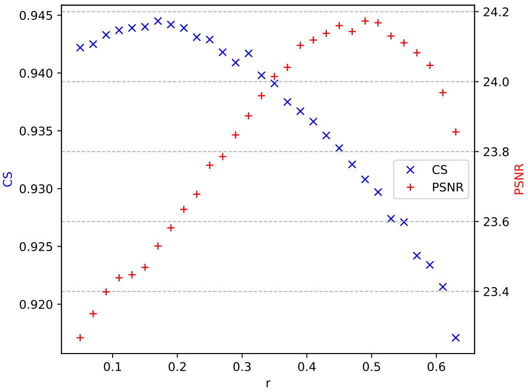


Figure: CS and PSNR as a function of r (sampling parameter). Higher values of r produce smoother images (with higher values of PSNR) but can decrease the value of CS

Results

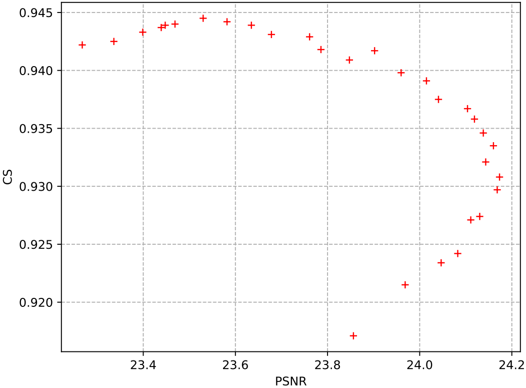


Figure: Cross-plot between CS and PSNR. The correlation coefficient between CS and PSNR is -0.6591, implying that higher values of PSNR do not always result in higher values of CS.

Results

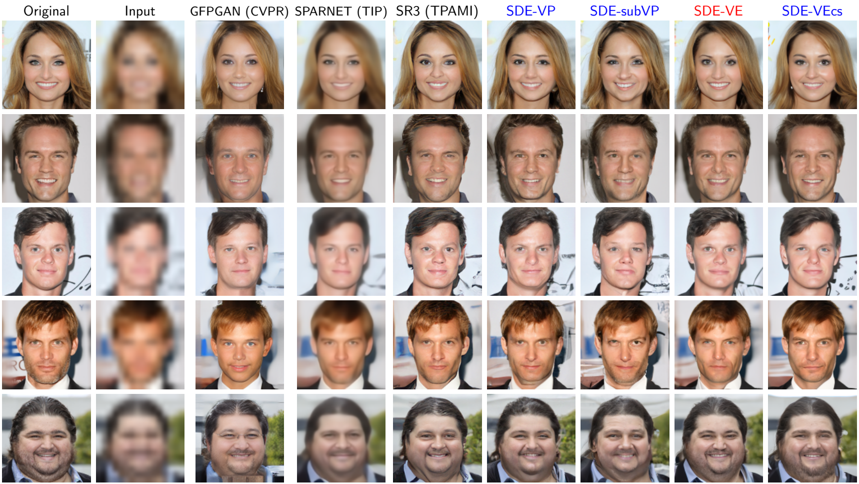


Figure: Super-resolution results. Our methods are shown in red (the best) and blue. SDE-VE provides more natural and detailed images than other methods.

Results

Model	PSNR ↑	SSIM ↑	CONSISTENCY ↓	CS ↑
GFP-GAN [2] (CVPR)	21.5326 ± 1.5273	0.6006 ± 0.0709	37.2256 ± 12.4622	0.8689 ± 0.0581
SPARNet [3] (TIP)	24.3686 ± 1.7844	0.7223 ± 0.0679	13.6512 ± 4.9063	0.9307 ± 0.0301
SR3 [4] (TPAMI)	22.9581 ± 1.8370	0.6605 ± 0.0758	1.3715 ± 0.7904	0.9370 ± 0.0244
SDE-VP	22.7171 ± 1.8107	0.6448 ± 0.0787	0.1074 ± 0.0592	0.9330 ± 0.0262
SDE-subVP	22.6455 ± 1.8047	0.6428 ± 0.0797	0.1433 ± 0.1212	0.9300 ± 0.0261
SDE-VE	23.5101 ± 1.9492	0.6879 ± 0.0797	0.0454 ± 0.0357	0.9443 ± 0.0222

Table: PSNR, SSIM, Consistency and CS on $16 \times 16 \rightarrow 128 \times 128$ face super-resolution. The best result for CS is highlighted with red.

Results

SCface - Surveillance Cameras Face Database - `https://www.scface.org/`

Original

Original'

SR

Reference

Results

SCface - Surveillance Cameras Face Database - `https://www.scface.org/`

Original

Original'

SR

Reference

- Increase in CS from 0.6415 ± 0.0633 to 0.6983 ± 0.0827.

Conclusions

Conclusions and future works

- SDEs can be successfully applied to SR problems.

Conclusions and future works

- SDEs can be successfully applied to SR problems.
- It was demonstrated the superior results of SDE-VE SR algorithm when using CS and Consistency metrics and qualitative analysis.

Conclusions and future works

- SDEs can be successfully applied to SR problems.
- It was demonstrated the superior results of SDE-VE SR algorithm when using CS and Consistency metrics and qualitative analysis.
- The SDE-VE SR algorithm has potential to be used for the recognition tasks in surveillance scenarios due to its noise removal property.

Conclusions and future works

- SDEs can be successfully applied to SR problems.
- It was demonstrated the superior results of SDE-VE SR algorithm when using CS and Consistency metrics and qualitative analysis.
- The SDE-VE SR algorithm has potential to be used for the recognition tasks in surveillance scenarios due to its noise removal property.
- The influence of the image smoothness on PSNR values and recognition accuracy will be further explored in future works.

Conclusions and future works

- SDEs can be successfully applied to SR problems.
- It was demonstrated the superior results of SDE-VE SR algorithm when using CS and Consistency metrics and qualitative analysis.
- The SDE-VE SR algorithm has potential to be used for the recognition tasks in surveillance scenarios due to its noise removal property.
- The influence of the image smoothness on PSNR values and recognition accuracy will be further explored in future works.
- Diffusion models and SDE based algorithms are computationally expensive.

References I

[1] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, pages 1–36, May 2021.

[2] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9164–9174, 2021.

[3] Chaofeng Chen, Dihong Gong, Hao Wang, Zhifeng Li, and Kwan-Yee K Wong. Learning spatial attention for face super-resolution. *IEEE Transactions on Image Processing*, 30:1219–1231, 2020.

[4] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Thank you!