Photo-Realistic Single Image Super-Resolution by Attention Diffusion Method

Xudong Wang[1+], Tao Liu[2 +], Zehang Chen[3+], Qingyuan Zeng[4 *], Tao Zeng[5*]

[*]Institute of Artificial Intelligence , Xiamen University
[+]School of Information Science and Technology, Xiamen University
Student-number-1: 31520221154224, Student-number-2: 31520221154215,
Student-number-3: 31520221154198, Student-number-4: 36920221153145,
Student-number-5: 36920221153146

Abstract

The main task of Single image superresolution (SISR) is to restore a given low resolution image into a high resolution image through a specific algorithm. When a low resolution image is given, it usually corresponds to many high resolution images, so it can be said that the reconstruction process is very uncertain. A review of previous super resolution methods shows that there are all kinds of problems, such as training instability and mode collapse in Gan-driven methods, so we propose a single image super resolution attention diffusion probability model (SRADM). SRADM takes low resolution image as input, converts Gaussian noise gradually into super resolution image through Markov chain, which can effectively enhance the quality of high resolution image. Here, we also introduce attention mechanism and residual prediction, which can effectively improve the model performance. We do a lot of experiments on DIV2K and other data sets, and finally can get a variety of super resolution results, and the model is easy to train, the effect is good.

Introduction

Single Image Super-resolution Reconstruction (SISR) aims to reconstruct high resolution (HR) images with clear detailed features from a given low resolution (LR) image. Image super resolution was first proposed by Harris (HARRIS 1964) in the 1960s, aiming to reconstruct a high resolution image from a low resolution image. (TSAI 1984) used multiple low-resolution images to restore high-resolution images in 1984. With the research and development of machine learning technology, Freeman et al. (FREEMAN, PASZTOR, and CARMICHAEL 2000) applied machine learning method to the field of image super resolution for the first time in 2000. Subsequently, a large number of super resolution methods based on machine learning emerged, such as the method based on neighborhood embedding (CHANG, YYEUNG, and XIONG 2004), the method based on sparse representation (YANG, WRIGHT, and HUANG 2008) and the method based on local linear regression (TIMOFTE, DE, and VAN 2013). However, most of these methods use the underlying features of images for super resolution reconstruction, and the expression ability of features is limited, which limits the reconstruction effect to a large extent.

Deep learning method can adaptively learn deep features from the training set and has been widely used in the field of image super resolution in recent years. In 2014, Dong et al. (DONG, LOY, and HE 2015) used Convolutional neural networks (CNNs) to directly learn the nonlinear mapping relationship between low-resolution images and high-resolution images, and the reconstruction effect has been greatly improved compared with the traditional methods. Since then, researchers have proposed a large number of deep learning-based superresolution network models, such as the use of residual learning and residual modules to construct deep super-resolution network models (LIM, SON, and KIM 2017), recursive structures (KIM, KWOn, and Mu 2016) and dense connections (ZHANG, TIAN, and KONG 2018). However, the gan driven method (Cheon et al. 2018) (Kim et al. 2019) also combines content loss and antagonistic loss well, so that the super resolution image with higher quality can be obtained. However, the gan driven method is also prone to problems such as mode collapse, and the generated super resolution image does not have the characteristics of diversity, and the model is not easy to converge in the training process.

At present, the diffusion model has a very powerful function in image generation (Ho, Jain, and Abbeel 2020). The diffusion model uses Markov chain to continuously add noise to the original data in the diffusion process to obtain Gaussian noise, and then recovers the original data from the Gaussian noise through the reverse diffusion process. The diffusion model is trained by optimizing a variable of the lower bound of variational, which can effectively solve the problem of mode collapse in gan driven method. The Attention Mechanism (Fei et al. 2017), as an effective means of feature screening and enhancement, has been widely applied in many fields of deep learning. It can be used to emphasize or select important information of the object and suppress some irrelevant details. It can bring great performance improvement to image processing, and we apply it in our super resolution diffusion model.

In this paper, we propose a single image Super Resolution Attention Diffusion probability Model(SRADM), which can solve the problem of pattern collapse in gan method and effectively improve the performance of the model by using the attention mechanism. First, SRADM iteratively recovers the original data in the process of diffusion and reverse

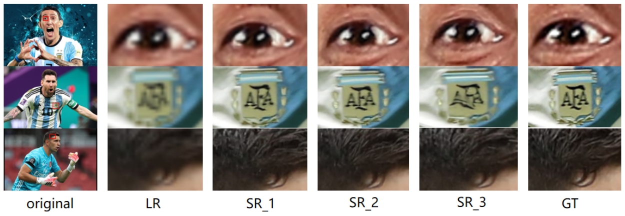

Figure 1: Selected some results from our 2040×1356 model(4×prediction). The SR predictions is richer in detail than LR image. As you can see from 3rd to 5th column, The resulting images are varied in texture detail.

diffusion, and rebuilds a given single low-resolution image into a high-resolution object. At the same time, we introduce residual prediction and attention mechanism into the model. Residual prediction can accelerate the model convergence and make the training effect more stable, while attention mechanism can enhance the model performance and improve the quality of super resolution image. Both of them enable SRADM to recover image details better. Our SRADM has the following two advantages: 1) Firstly, super resolution can obtain high quality images, and the image is also characterized by diversity, which effectively solves the disadvantages of gan method; 2) The training occupies less space, is convenient for training, and the training process is stable and efficient, and the training speed is fast.

Finally, our experiments on DIV2K(Timofte et al. 2018) and other data sets proved that: 1) our SRADM model can reconstruct different super resolution images when a single image is input, and the reconstructed images of the model are more diverse; 2) compared with some models, the training is faster and the number of parameters is less.

Related Works

Single Image Super-Resolution

In recent years, deep learning methods have been increasingly applied to the super resolution of single images. Firstly, SRCNN (DONG, LOY, and HE 2015) sets a precedent for the end-to-end mapping between LR and HR images. SRCNN first undersamples the images to obtain LR images, and then enlarging the image to the target resolution by using bicubic interpolation, and then using three convolution layers of different sizes. Feature extraction was completed, nonlinear mapping between LR-HR image pairs was fitted, and output results of the network model were reconstructed. Finally, the final HR image was obtained. Then, FSRCNN (Chao Dong 2019) improved SRCNN: 1) Directly used LR image as input, reducing the feature dimension; 2) Using a smaller filter than SRCNN, the network structure is deepened; 3) Adopt the back-end up-sampling hyperdivision framework, and add deconvolution layer at the end of the network to enlarge the image to the target resolution.

After all kinds of convolutional neural network-based super resolution algorithms, super resolution algorithms based on generative adversarial network emerge, which have more prominent effects in image reconstruction effect, network computation amount and operation speed compared with the former. SRGAN (Ledig et al. 2017) algorithm applies the generator network and discriminator network confrontation training to super resolution image reconstruction for the first time. It uses the generator to generate HR image, the discriminator to discriminate the reconstructed HR image and the original HR image, and reversely optimizes the generator network and discriminator network. At the same time, ”perceptual loss” is used to replace the traditional MSE loss function to enhance the restoration of image details and ensure the high fidelity and high quality of the reconstructed image. While ESRGAN (Wang et al. 2018) enhances the performance of SRGAN, improves the generalization ability of network, uses residual scaling to accelerate the training speed of deep network and reduces the number of network operation parameters, so that the reconstructed HR image has richer texture features and the color brightness is closer to the original HR image.

In terms of data sets, image data sets of super resolution image reconstruction used for deep learning involve many fields, covering people, animals and plants, buildings, natural landscapes, etc., and many open source data sets differ greatly in external conditions (resolution size, number of sheets, format, etc.) and internal conditions (content, style, texture, etc.) of images.

Diffusion models

Diffusion probability model (Sohldckstein, Eric A Weiss, and Ganguli 2015) is a kind of generation model. The diffusion model uses Markov chain to gradually apply noise to the image in the forward stage until the image is destroyed into complete Gaussian noise, and then learns the process of restoring the image from Gaussian noise to the original image in the reverse stage. The diffusion model helps to enhance the diversity of the generated results and the quality of the generated results is high. Diffusion models have not been widely used recently in the field of image reconstruction. Our SRADM model is capable of producing diverse and high quality image results.

Attention mechanism

Attention mechanism was first proposed in the field of visual images, and then it was used in image classification (Mnih et al. 2014) and machine translation task (Bahdanau, Cho, and Bengio 2014), and then it was widely used in various NLP tasks based on neural network models such as RNN or CNN. The attention mechanism has the characteristics of fewer parameters, fast speed and good effect, which can significantly improve the quality of image generation. In this paper, the attention mechanism is introduced into the model, which further improves the effect of image generation and improves the performance of the model.

The proposed method

Diffusion Model

Diffusion models are inspired by non-equilibrium thermodynamics. They define a Markov chain of diffusion steps to slowly add random noise to the data, and then learn to

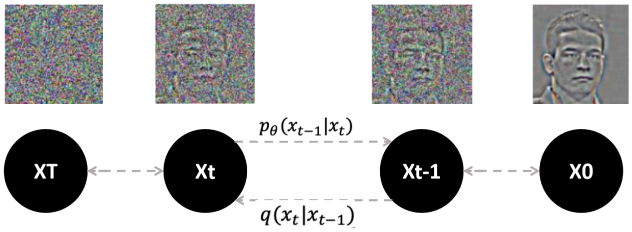

Figure 2: Overview of two processes in SRADM

reverse the diffusion process to construct the desired data samples from the noise. Unlike VAEs or flow models, diffusion models are learned through a fixed process and the latent variables are high-dimensional (same as the original data). In this section, we will briefly introduce it.

There has been previous work on similar diffusion models, including diffusion probabilistic models (Sohldckstein, Eric A Weiss, and Ganguli 2015), noise-conditioned score network (Song and Ermon 2020), and denoising diffusion probabilistic models (Ho, Jain, and Abbeel 2020).

We define a forward diffusion process in which we add small amount of Gaussian noise to the sample in T steps and transform the input image into pure Gaussian noise . And our model is responsible for restoring back to image . In this way, the diffusion model is actually very similar to GAN, which generates a picture with a given noise , but it should be emphasized that the noise and picture are of the same dimension. The diffusion model include two processes: forward diffusion process and reverse diffusion process.

The posterior $q(x_1,\cdots,x_t|x_0)$, called the diffusion process, transforms the data distribution $q(x_0)$ into a latent variable distribution $q(x_t)$, fixed as a Markov chain that gradually Gaussian noise is added to the data according to the variance table β_1, β_2.

$$q\left(x_1, \cdots, x_T \mid x_0\right):=\prod_{t=1}^T q\left(x_t \mid x_{t-1}\right) \tag{1}$$

$$q\left(x_t \mid x_{t-1}\right):=\mathcal{N}\left(x_t ; \sqrt{1-\beta_t} x_{t-1}, \beta_t \mathbf{I}\right) \tag{2}$$

where β_t is a small positive number that can be treated as a constant hyper-parameter. Setting $\overline(\alpha_t) := 1-\beta_t$, $\bar{\alpha}_t := \prod_{s=1}^t \alpha_s$, as the diffusion process allows sampling x_t in closed form at any time step t:

$$q\left(x_t \mid x_0\right)=\mathcal{N}\left(x_t ; \sqrt{\bar{\alpha}_t} x_0,\left(1-\bar{\alpha}_t\right) \mathbf{I}\right) \tag{3}$$

Then can be further reparameterized as

$$x_t\left(x_0, \epsilon\right)=\sqrt{\bar{\alpha}_t} x_0+\sqrt{1-\bar{\alpha}_t} \epsilon, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{4}$$

The inverse process transforms the latent variable distribution p$_\theta\left(x_t\right)$ into a data distribution p$_\theta\left(x_0\right)$ parameterized by θ. It is defined by a Markov chain with a learned Gaussian transformation starting with $p\left(x_T\right)=\mathcal{N}\left(x_T ; \mathbf{0}, \mathbf{I}\right)$

$$p_\theta\left(x_0, \cdots, x_{T-1} \mid x_T\right):=\prod_{t=1}^T p_\theta\left(x_{t-1} \mid x_t\right) \tag{5}$$

$$p_\theta\left(x_{t-1} \mid x_t\right):=\mathcal{N}\left(x_{t-1} ; \mu_\theta\left(x_t, t\right), \sigma_\theta\left(x_t, t\right)^2 \mathbf{I}\right) \tag{6}$$

During the training phase, we maximize a variational lower bound (ELBO) on the negative log-likelihood and introduce KL divergence and variance reduction[3].

$$\begin{aligned} & E\left[-\log p_\theta\left(x_0\right)\right] \leq L:=E_q[\underbrace{D_{KL}\left(q\left(x_T \mid x_0\right) \| p\left(x_T\right)\right)}_{L_T} \\ & +\sum\nolimits_{t>1} \underbrace{D_{KL}\left(q\left(x_{t-1} \mid x_t, x_0\right) \| p_\theta\left(x_{t-1} \mid x_t\right)\right)}_{L_0} \\ & \underbrace{-\log p_\theta\left(x_0 \mid x_1\right)}_{L_0}] \end{aligned} \tag{7}$$

This shift requires a direct comparison of $p_\theta\left(x_{t-1} \mid x_t\right)$ and its corresponding diffusion process posterior. Setting $\tilde{\mu}_t\left(x_t, x_0\right):=\frac{\sqrt{\bar{\alpha}_{t-1}} \beta_t}{1-\bar{\alpha}_t} x_0+\frac{\sqrt{\alpha_t}\left(1-\bar{\alpha}_{t-1}\right)}{1-\bar{\alpha}_t} x_t$, we have equivalent with:

$$q\left(x_{t-1} \mid x_t, x_0\right)=\mathcal{N}\left(x_{t-1} ; \tilde{\mu}_t\left(x_t, x_0\right), \tilde{\beta}_t \mathbf{I}\right) \tag{8}$$

Eq.(3),(5),(6) and (8) ensure all KL divergences in Eq.(7) is a comparison between Gaussians, where $\sigma_\theta^2=\tilde{\beta}_t=\frac{1-\bar{\alpha} t-1}{1-\bar{\alpha}_t} \beta_t for\, t>1, \tilde{\beta}_1=\beta_1$, and constant C, we have:

$$L_{t-1}=E_q\left[\frac{1}{2 \sigma_t^2}\left\|\tilde{\mu}_t\left(x_t, x_0\right)-\mu_\theta\left(x_t, t\right)\right\|^2\right]+C \tag{9}$$

For simplicity, the training procedure minimizes the variant ELBO with and t as input:

$$\min_\theta L_{t-1}(\theta)=E_{x_0, \epsilon, t}\left[\left\|\epsilon-\epsilon_\theta\left(\sqrt{\bar{\alpha}_t} x_0+\sqrt{1-\bar{\alpha}_t} \epsilon, t\right)\right\|^2\right] \tag{10}$$

where ε_θ is the noise predictor.

In inference, we first sample an $x_T \sim \mathcal{N}\left(x_T ; \mathbf{0}, \mathbf{I}\right)$ and sample $x_{t-1} \sim p_\theta\left(x_{t-1} \mid x_t\right)$ according to Eq.(5),(6), where

$$\mu_\theta\left(x_t, t\right):=\frac{1}{\sqrt{\alpha_t}}\left(x_t-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta\left(x_t, t\right)\right) \tag{11}$$

$$\sigma_\theta\left(x_t, t\right):=\tilde{\beta}_t^{\frac{1}{2}}, t \in\{T, T-1, \ldots, 1\} \tag{12}$$

Self attention

$$z_i=\sum_{j=1}^n SoftMax\left(\left(x_i W^Q\right)\left(x_j W^K\right)^T\right)\left(x_j W^V\right) \tag{13}$$

Transformer (A, N, and N 2017) was originally proposed for the NLP field and has been very successful in the NLP

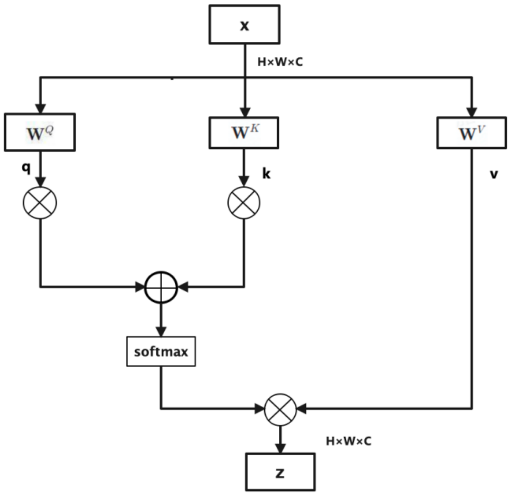

Figure 3: Self-attention mechanism process

field. Vit(A, L, and A 2020) tries to apply Transformer to the CV field.Self-attention maps a query and a set of key-value pairs to an output. More specifically, for an input sequence, such as an embedding of words or image patches, $\mathrm{x} = (x_1, \ldots, x_n)$ of n elements where $\mathbf{x}_i \in R^{d_x}$, the self-attention module computes an output sequence $\mathrm{z} = (z_1, \ldots, z_n)$ of the same length, where $\mathbf{z}_i \in R^{d_z}$. Each output element z_i is computed as a weighted sum of linearly transformed input elements: The process is shown in the figure 1. The calculation process of Self Attention is as follows:

Among them, $W^Q, W^K, W^V \in \mathrm{R}^{\mathrm{d_x} \times \mathrm{d_z}}$ is a parameterized matrix, W^Q, W^K, W^V are all learnable parameter matrices, and x represents Self Attention The vector in the middle window (windows), z represents the output vector of self attention, self-attetion divides the feature map into n non-overlapping windows according to the window size, where i represents the i-th window (windows), j represents The j-th window (windows).

SRADM

As shown in Figure 2, SRADM is built on the T-step diffusion model, which contains two processes: the diffusion process and the inverse process. Instead of predicting the HR image directly, we apply residual prediction to predict the difference between the HR image x_H and the upsampled LR image $up(x_L)$, denoting the difference as the input residual image x_0. The diffusion process transforms x_0 into an underlying x_T in a Gaussian distribution by gradually adding the Gaussian noise ε implicit in Eq.(4). According to the equation. (5),(6),(11) and (12), the inverse process is determined by ε_θ, which is a conditional noise predictor based on RRDB (Wang et al. 2018) Low-resolution encoder (LR encoder for short) D. The reverse process transforms the latent variable x_T into a residual image x_r by iteratively denoising in finite steps T using a conditional noise predictor ε_θ, encoded from the LR image by the hidden state conditional LR encoder D. The SR image is reconstructed by adding the resulting residual image x_r to the upsampled LR image $up(x_L)$. Therefore, the goal of ε_θ is to predict the noise ε.

Conditional Noise Predictor

The conditional noise predictor ε_θ predicts the noise added at each time step of the diffusion process conditioned on the LR image information, according to Eq.(10),(11) and (12). As shown in Fig.2, we take U-Net as the main body, and take the output of the 3-channel x_t, diffusion time step $t \in \{1, 2, \ldots, T-1, T\}$ LR encoder as input. First, x_t is converted to a hidden layer by a 2D convolutional block consisting of a 2D convolutional layer and a Mish activation layer (Misra 2019). The LR information is then fused with the hidden 2D convolutional block output. We use Transformer sinusoidal position encoding [Vaswani et al., 2017] to convert time step t into time step embedding t_e. The last outputs hidden and t_e are then fed sequentially into the contraction path, an intermediate step, and the dilation path. Both the shrinkage path and the expansion path consist of four steps, each of which sequentially applies two residual blocks and a downsampling/upsampling layer. To reduce the model size, we only double the channel size in the second and fourth shrinkage steps and halve the spatial size of the feature maps in each shrinkage step. The downsampling layer in the shrinking path is a two-step 2D convolution, and the upsampling layer in the dilation path is a 2D transposed convolution. The intermediate step consists of two residual blocks, inserted between the shrinking path and the expanding path. Furthermore, the input connections of each expansion step come from the corresponding feature maps of the contraction path.

Finally, a 2D convolutional block is applied to generate $\hat{\epsilon}$ in time step t-1 as prediction noise, which is then used to recover x_{T-1} according to Eq.(5),(6),(11) and (12). Our conditional noise predictor is easy to train and stable due to multi-scale skip connections. Furthermore, it combines local and global information through contraction and expansion paths.

LR Encoder

The LR encoder encodes the LR information x_e, which is added to each reverse step to steer the generation to the corresponding HR space. In this paper, we choose to follow the RRDB architecture of SRFlow (Lugmayr et al. 2020), which employs a residual-in-residual structure and multiple dense skip connections without batch normalization layers. We discard the last convolutional layer of the RRDB architecture, since we do not target specific SR results but hide LR image information.

Experiments

Datasets

SRADM is trained and evaluated on face SR (8x) and general SR (4x) tasks. For face SR, we use the Celeb-Faces Attributes Dataset (CelebA) (Liu et al. 2015), a large face attribute dataset containing over 200,000 celebrity images. The images in this dataset cover large pose changes and background clutter. In this paper, we train and evaluate his

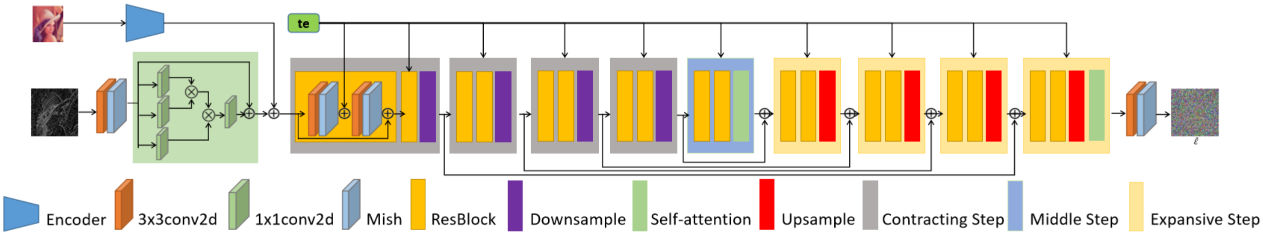

Figure 4: Our proposed method framework

5,000 images from the test split of SRFlow using the entire training set consisting of 162,770 images. We crop the aligned patches intensively using standard MATLAB bicubic kernels and resize them to 160 × 160 as HR ground truth. Downsample the HR image using a bicubic kernel to obtain the corresponding LR image. For ProgFSR(Kim et al. 2019), We use the performed double -line cores introduced in its original papers for fair comparison.

Training

Training and evaluation First of all, we used the L1 loss to iterate for 100K and conducted a preview for the LR encoder D to achieve efficiency. Training and use of conditional noise predictors. (6) As a loss term, Adam (Kingma and Ba 2014) as a optimizer, batch size 16 and learning rate, half of the steps of 100K. The entire SRADM takes about 34/45 hours (300K/400K steps) and trains on a 11GB memory GeForce RTX 2080TI, which is CELEBA/DIV2K.

Performance

In this subsection, we evaluate SRADM by comparing with serveral SR methods on CelebA and DIV2K.The detailed configuration of the baseline model can be found in their original paper.

Table 1: Result for 8×SR of faces on CelebA.

Method	PSNR	SSIM	LPIPS	LR-PSNR
Bicubic	23.37	0.65	0.483	34.65
RRDB	26.89	0.78	0.220	48.00
ESRGAN	23.25	0.66	0.115	39.91
SRADM	25.35	0.73	0.105	52.32

As shown in Table 1, for most evaluation indicators of SR (PSNR, SSIM, and LR-PSNR) and comparable LPIPS, the quantitative results of SRADM are comparable to those of previous methods, which indicates that our method is feasible. Figure 5 indicates that SRADM well balanced degree of sharpness and natural, and produced a strong consistency with the LR images.Compared with GAN, our model can generate exquisite details and overcome the defects of GAN model in diversity. Through multiple sampling, the model can predict a variety of high-frequency information, which forms numerous and exquisite details of the image. Moreover, compared with SRfow, the structure of our model is relatively simple, which only contains small parameters and has low requirements on hardware. Therefore, better results can be trained in a relatively short time.After about 15 hours of running on the 3090 device, the model was able to converge.

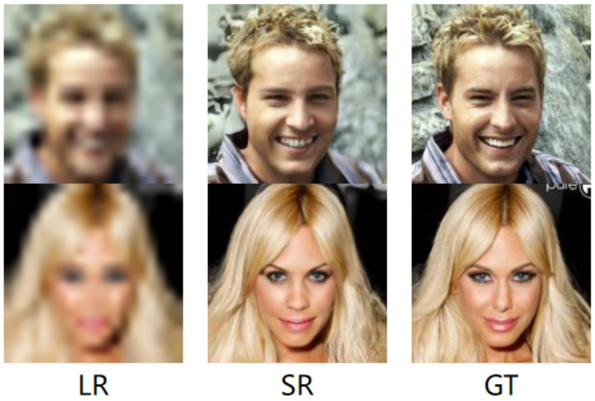

Figure 5: Some results(8×) in CelebA. SRADM predicts rich details and keep consistency with the ground true.

We also evaluate SRADM on DIV2K (4×) and compare it with RRDB, ESRGAN, RankSRGAN. As shown in Table 2, for most of the evaluation indicators (PSNR, SSIM, and LR-PSNR) and comparable LPIPS, SRADM quantified results better than previous methods, demonstrating the effectiveness and great potential of our approach. Figure 5 shows that SRADM strikes a good balance between sharpness and naturalness and produces a strong consistency with LR images.

Table 2: Results for 4×SR for DIV2K Dataset.

Method	PSNR	SSIM	LPIPS	LR-PSNR
Bicubic	26.71	0.77	0.410	38.71
RRDB	28.99	0.83	0.270	54.91
RankSRGAN	26.55	0.75	0.128	42.33
ESRGAN	23.25	0.66	0.115	39.91
SRADM	27.33	0.79	0.137	55.43

In order to further test the generalization ability of the model on non-SR data sets, we collected a considerable

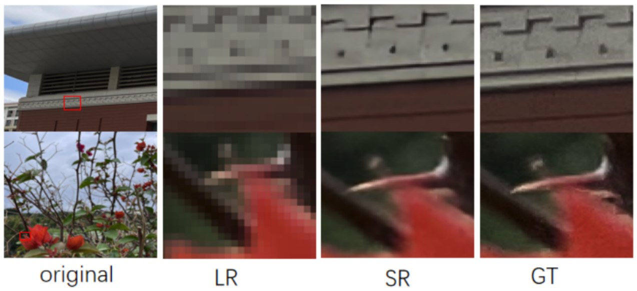

Figure 6: The successful samples of SRADM. Model generate rich details than Low-resolution images and and maintain consistency with the ground truth.

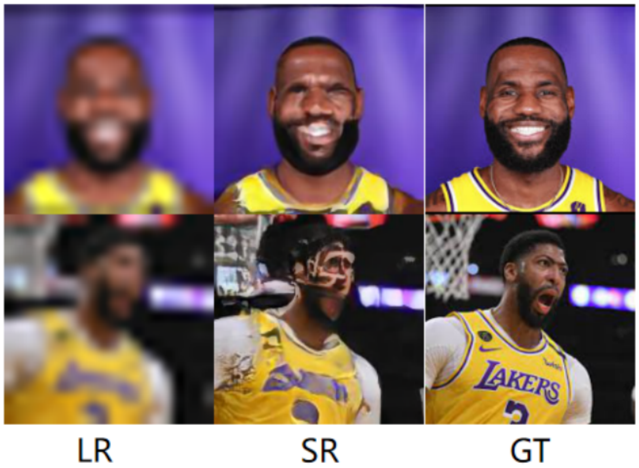

Figure 7: Some example of failure.The details generated do not meet expectations.

number of non-SR images for testing. Multiple experiments show that the proposed model can still achieve satisfactory results in terms of general data. Although not every data generation results live up to expectations, the percentage of failures is within acceptable limits.Figure 6 exhibits some successful samples of our model, as the result shows, The images generated by the model not only retain the semantic information of LR image, but also generate rich and delicate details.

Due to the randomness of sampling, the SR images obtained are diverse and close to the GT images in quality. In addition, due to the randomness of sampling, the occurrence of some events is bound to fail to meet expectations, as shown in Figure 7.Not only that, we also noticed that when the quality of the low resolution image itself is poor, the resulting high resolution image does not improve much as the figure 8 shows.

Conclusion

In this article, we proposed SRADM. Our work uses the Malcov chain to convey the HR image to the incubation period with a simple distribution standard, and then perform SR prediction during the reverse process. Essence In order

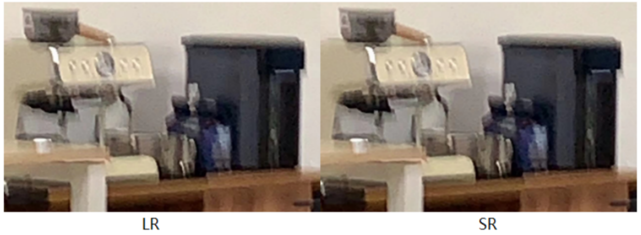

Figure 8: Prediction from poor quality image.

to speed up integration and stable training, SRADM introduced residual predictions. And add self -attention mechanisms to make the model more concerned about important regional characteristics. We have conducted extensive experiments on facial and general datasets that SRADM can generate diverse and realistic SR images and avoid excessive smoothness and pattern collapse in PSNR -oriented methods and GAN drive methods, respectively. In addition, SRADM is trained with a small amount of footprint without extra discrimination. In addition, SRADM allows flexible image manipulation, including potential space interpolation and content fusion.

In the future, we will further improve the performance of diffused SISR models and speed up inference. We will also expand work to more image recovery tasks (for example, image Denoising, Debluring and Dehazing) to verify the potential of the diffusion model in the image recovery domain.

Acknowledgments

References

A, D.; L, B.; and A, K. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv2010.11929*.

A, V.; N, S.; and N, P. 2017. Attention is all you need. *Conference and Workshop on Neural Information Processing Systems(NIPS)*, 30.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv:1409.0473*.

CHANG, H.; YYEUNG, D.; and XIONG, Y. 2004. Superresolution through neighbor embedding. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 275–282.

Chao Dong, X. T., Chen Change Loy. 2019. Accelerating the Super-Resolution Convolutional Neural Network. *arXiv: 1608.00367*.

Cheon, M.; Kim, J.-H.; Choi, J.-H.; and Lee, J.-S. 2018. Generative adversarial network-based image superresolution using perceptual content losses. *Proceedings of European Conference on Computer Vision(ECCV)*, 0–0.

DONG, C.; LOY, C. C.; and HE, K. 2015. Image superresolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2): 295–307.

Fei, W.; Jiang, M.; Chen, Q.; and Yang, S. 2017. Residual Attention Network for Image Classification. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR).*

FREEMAN, W. T.; PASZTOR, E. C.; and CARMICHAEL, O. T. 2000. Learning low-level vision. *International Journal of Computer Vision*, 40(1): 24–47.

HARRIS, J. 1964. Diffraction and Resolving Power. *Journal of the Optical Society of America*, 54(7): 931–936.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Conference and Workshop on Neural Information Processing Systems(NIPS)*, 33: 6840–6851.

Kim, D.; Kim, M.; Kwon, G.; and Kim, D.-S. 2019. Progressive face super-resolution via attention to facial landmark. *arXiv*, 0–0.

KIM, J.; KWOn, L. J.; and Mu, L. K. 2016. Deeply-recursive convolutional network for image super-resolution. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 1637–1645.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *ArXiv1412.6980.*

Ledig, C.; Theis, L.; Huszar, F.; Caballero, J.; and Cunningham, A. 2017. Photo-realistic single image super-resolution using a generative adversarial network. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR).*

LIM, B.; SON, S.; and KIM, H. 2017. Enhanced deep residual networks for single image super-resolution. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 136–144.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. *Proceedings of the IEEE international conference on computer vision(ICCV)*, 3730–3738.

Lugmayr, A.; Danelljan, M.; Gool, L. V.; and Timofte, R. 2020. Srflow: Learning the super-resolution space with normalizing flow. *Proceedings of European Conference on Computer Vision(ECCV)*, 715–732.

Misra, D. 2019. Mish: A self regularized non-monotonic activation function. *ArXiv1908.08681.*

Mnih, V.; Heess, N.; Graves, A.; and Kavukcuoglu, K. 2014. Recurrent Models of Visual Attention. *arXiv:1406.6247.*

Sohldckstein, J.; Eric A Weiss, N. M.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. *arXiv:1503.03585.*

Song, Y.; and Ermon, S. 2020. Generative Modeling by Estimating Gradients of the Data Distribution. *arXiv.1907.05600.*

TIMOFTE, R.; DE, S. V.; and VAN, L. G. 2013. Anchored neighborhood regression for fast example-based super-resolution. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 1920–1927.

Timofte, R.; Gu, S.; Wu, J.; and Gool, L. V. 2018. NTIRE 2018 challenge on single image super-resolution: methods and results. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR) workshop*.

TSAI, R. 1984. Multi-frame image restoration and registration. *Advance Computer Visual and Image Processing*, 1: 317–339.

Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; and Loy, C. C. 2018. Enhanced superresolution generative adversarial networks. *Proceedings of European Conference on Computer Vision(ECCV).*

YANG, J.; WRIGHT, J.; and HUANG, T. 2008. Image super-resolution as sparse representation of raw image patches. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–8.

ZHANG, Y.; TIAN, Y.; and KONG, Y. 2018. Residual dense network for image super-resolution. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2472–2481.