ResDiff: Combining CNN and Diffusion Model for Image Super-Resolution

Shuyao Shang [1], Zhengyang Shan [1], Guangxing Liu [1], LunQian Wang [2], XingHua Wang [2], Zekai Zhang [3], Jinglin Zhang* [1]

[1] Shandong University
[2] Linyi University
[3] Qilu University of Technology
202000800098@mail.sdu.edu.cn, 202000800128@mail.sdu.edu.cn, 202000800141@mail.sdu.edu.cn,
210854002032@lyu.edu.cn, 210854002052@lyu.edu.cn, 10431210745@stu.qlu.edu.cn, jinglin.zhang@sdu.edu.cn

Abstract

Adapting the Diffusion Probabilistic Model (DPM) for direct image super-resolution is wasteful, given that a simple Convolutional Neural Network (CNN) can recover the main low-frequency content. Therefore, we present **ResDiff**, a novel **Diff**usion Probabilistic Model based on **Res**idual structure for Single Image Super-Resolution (SISR). ResDiff utilizes a combination of a CNN, which restores primary low-frequency components, and a DPM, which predicts the residual between the ground-truth image and the CNN-predicted image. In contrast to the common diffusion-based methods that directly use LR space to guide the noise towards HR space, ResDiff utilizes the CNN's initial prediction to direct the noise towards the residual space between HR space and CNN-predicted space, which not only accelerates the generation process but also acquires superior sample quality. Additionally, a frequency-domain-based loss function for CNN is introduced to facilitate its restoration, and a frequency-domain guided diffusion is designed for DPM on behalf of predicting high-frequency details. The extensive experiments on multiple benchmark datasets demonstrate that ResDiff outperforms previous diffusion-based methods in terms of shorter model convergence time, superior generation quality, and more diverse samples.

Introduction

Single Image Super-Resolution (SISR) is a difficult task in computer vision, which aims to recover high-resolution (HR) images from their low-resolution (LR) counterparts. During image degradation, the high-frequency components are lost, and multiple HR images could produce the same LR image, making this task ill-posed. After Generative Adversarial Networks(GAN) (Goodfellow et al. 2014) was proposed, the main generative-model-based SISR methods are GAN-driven. However, GAN-based methods are hard to train and prone to fall into pattern collapse, causing a lack of diversity. Therefore, a superior generative model is required in the SISR task.

Diffusion Probabilistic Model (DPM) has already demonstrated impressive capabilities in image synthesis (Saharia et al. 2022a,b; Rombach et al. 2022; Ramesh et al. 2022) and image restoration (Choi et al. 2021; Kawar et al. 2022; Wang, Yu, and Zhang 2023). It has also shown promising prospects in SISR tasks (Saharia et al. 2022c; Li et al. 2022). However, current Diffusion-based methods for SISR, such as SR3(Saharia et al. 2022c), generate HR images directly from random noise, and LR images are only used as conditional input to the diffusion process (Fig.2 (a)). Consequently, the diffusion model needs to recover both the high and low-frequency contents of the image, which not only prolongs the convergence time but also inhibits the model from focusing on the fine-grained information, potentially missing texture details. Li et al.(Li et al. 2022) had taken this into account but employed only a bilinear interpolation for the initial prediction, which, compared to CNN, failed to restore sufficiently low-frequency contents and was incapable of generating any high-frequency components in the initial prediction (Fig.2 (b)). Similarly, whang et al.(Whang et al. 2022) designed a random-sampler and a deterministic-predictor to tackle this problem. However, there is no information interaction between the random-sampler and the deterministic-predictor, resulting in the latter not functioning to its full potential (Fig.2 (c)).

Inspired by the above (Li et al. 2022; Whang et al. 2022), we propose ResDiff, a residual-structure-based dif-

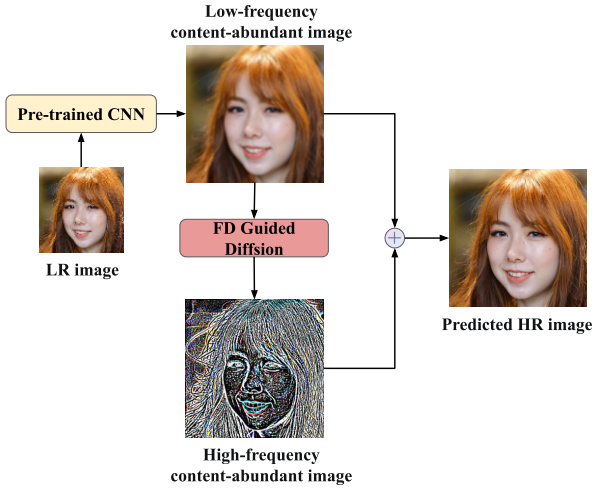

Figure 1: Overall struture of proposed ResDiff.

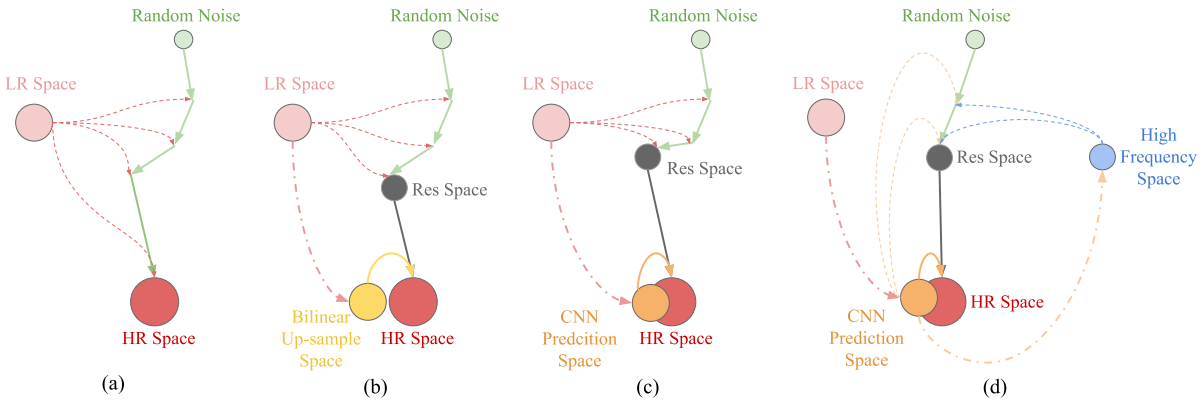

Figure 2: Comparison of different generation processes. In contrast to (a) (Saharia et al. 2022c), (b) (Li et al. 2022), (c) (Whang et al. 2022) where only LR Space is used to guide the generation, our ResDiff (d) makes full utilization of CNN Prediction Space and High-Frequency Space to guide a faster and better generation.

fusion model. Unlike (Li et al. 2022), ResDiff utilizes a CNN for initial prediction. And in contrast to (Whang et al. 2022), the CNN in ResDiff is pre-trained, thus capable of restoring the major low-frequency components and partial high-frequency components. The initial prediction of the CNN is adopted to guide the random noise towards the Res Space (i.e., the residual space between the Ground Truth image and the CNN predicted image). Compared to the methods that only use LR space as guidance, ResDiff can leverage additional information and generate richer high-frequency details. (Fig.2 (d)). Fig.1 presents the structure of ResDiff. The CNN used in ResDiff contains a limited number of parameters. Thus, two more loss functions are introduced to strengthen its recovery capabilities. To further enhance the generation quality, we design a Frequency Domain-guided Diffusion (FD-guided Diffusion) as shown in Fig.2 (d) where the high-frequency space also guides the generation process. FD-guided Diffusion consists of two novel modules. The first is a Frequency-Domain Information Splitter (FD Info Splitter) that separates high-frequency and low-frequency contents and performs adaptive denoising on the noisy image. The second is a high-frequency guided cross-attention module (HF-guided CA) that helps the diffusion model predict high-frequency details. The pseudo-code for sampling with ResDiff is as Alg.1.

Experiments on two face datasets (FFHQ and CelebA) and two general datasets (Div2k and Urban100) demonstrate that ResDiff not only accelerates the model's convergence speed but also generates more fine-grained images. To verify the generalization of our method, more experiments on different types of datasets (Bai et al. 2022) are given in the supplementary material.

Our contributions can be summarized as follows:

• **Shorter Convergence Time**: We have designed ResDiff, a residual structure-based diffusion model for the SISR task that leads to an apparent improvement in convergence speed compared to other diffusion-based methods.

• **Superior Generation Quality**: We have introduced FD-guided Diffusion to enhance the diffusion model's concentration on high-frequency details, resulting in superior generation quality.

• **More Diverse Output**: Experiments have demonstrated that ResDiff holds a lower perceptual-based evaluation value, indicating our method is capable of producing diverse samples.

Related Works

Generative-model-based methods have created great success in SISR, which can be classified into GAN-based(Ledig et al. 2017; Wang et al. 2018b; Mirchandani and Chordiya 2021; Wang et al. 2018a; Zhang et al. 2019), flow-based(Lugmayr et al. 2020; Liang et al. 2021), and fiffusion-based(Saharia et al. 2022c; Li et al. 2022) methods.

GAN-based methods Ledig et al.(Ledig et al. 2017) proposed SRGAN, which employs a perceptual loss function to generate high-quality images. Similarly, Kim et al.(Wang et al. 2018b) introduced ESRGAN, which adopted an enhanced super-resolution GAN and a superior loss function to improve the perceptual quality. GAN-based methods combine content losses with adversarial losses, allowing them to generate sharp edges and richer textures. However, they are prone to mode-collapse, which decreases diversity in the generated SR samples. Moreover, training GANs is challenging and may lead to unexpected artifacts in the generated image.

Flow-based methods Lugmayr et al.(Lugmayr et al. 2020) proposed SRFlow, which is a flow-based method that learns the conditional distribution of high-resolution images given their low-resolution counterparts, enabling high-quality image super-resolution with natural and diverse outputs. Flow-based methods map HR images to flow-space latents using an invertible encoder and connect the encoder and decoder with an invertible flow module, which avoids training instability but requires higher training costs and provides lower perceptual quality.

Algorithm 1: ResDiff Inference

Input: low-resolution image x_{LR} and pre-trained CNN
Parameter: μ_θ and Σ_θ same as in DDPM
Output: High-resolution image

x_{cnn} = CNN(x_{LR})
$x_T \sim \mathcal{N}(0, I)$
for $t = T : 1$ **do**
  $\epsilon \sim \mathcal{N}(0, I)$ if $t > 1$, else $\epsilon = 0$
  $x_{t-1} = \mu_\theta(x_t, t, x_{cnn}) + \sqrt{\Sigma_\theta(x_t, t, x_{cnn})}\,\epsilon$
end for
return $x_0 + x_{cnn}$

Diffusion-based methods Li et al.(Li et al. 2022) introduced SrDiff, the first diffusion-based model for SISR, demonstrating that using the diffusion model for SISR tasks is feasible and promising. Saharia et al. proposed Sr3 (Saharia et al. 2022c), which adapts Denoising Diffusion Probabilistic Models (DDPM) to perform SISR tasks, yielding a competitive perceptual-based evaluation value. Diffusion-based methods utilize a diffusion process that simulates noise reduction, resulting in sharper and more detailed images. However, a high computational cost is needed due to multiple forward and backward passes through the entire network during the training process. Our proposed ResDiff, though without improving the training speed of a single iteration, accelerates convergence, which can alleviate this issue from another perspective.

The Proposed ResDiff

Pre-trained CNN

To reduce additional training costs, we utilize a CNN with a reduced number of parameters to generate an initial prediction. This CNN aims to recover primary low-frequency components and partial high-frequency components, consequently facilitating the diffusion model's restoration of the more intricate high-frequency details. To ensure its generating capability, we are enlightened by (Deng et al. 2019; Dou, Tu, and Peng 2020) and introduce two more loss functions (Fig.3), namely $\mathcal{L}_{FFT}$ based on the Fast Fourier Transform (FFT) (Cooley and Tukey 1965) and $\mathcal{L}_{DWT}$ based on the Discrete Wavelet Transform (DWT) (Mallat and Hwang 1992), in addition to the original loss function.

The $\mathcal{L}_{FFT}$ can be defined as the mean square error(MSE) between the magnitudes of the FFT coefficients of the two images:

$$\mathcal{L}_{FFT} = \mathbb{E}[\left\| M - \hat{M} \right\|^2] \tag{1}$$

where M and $\hat{M}$ denote the frequency domain images obtained by performing FFT on the ground-truth image and the predicted image.

In a bid to enable the CNN to further recover partial high-frequency contents on top of recovering the primary low-frequency contents, we designed $\mathcal{L}_{DWT}$. Performing DWT on an image will decompose it into four sub-bands: low-low (LL), low-high (LH), high-low (HL), and high-high (HH).

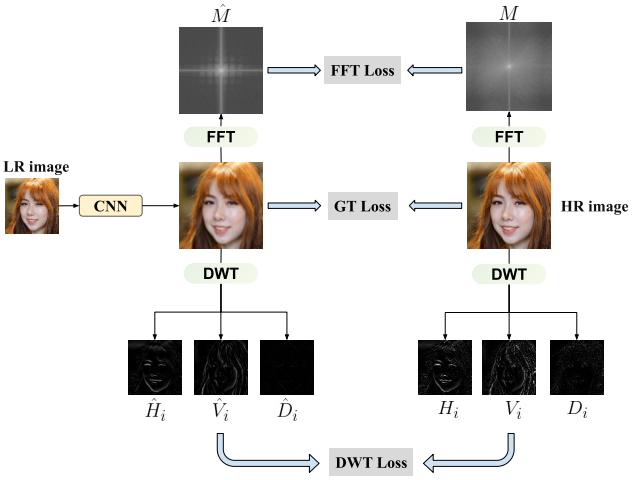

Figure 3: Depiction of the three loss functions utilized in CNN pre-training. A spatial domain loss (GT Loss) and two frequency domain losses (FFT Loss and DWT Loss) are computed.

LL sub-band contains the low-frequency content of the image, while the remaining three contain the high-frequency components of the image from horizontal, vertical, and diagonal directions, respectively. The LL sub-band can perform further similar decomposition to obtain multi-layer high-frequency components. As for $\mathcal{L}_{DWT}$, we extract the wavelet coefficients of the high-frequency bands H, V, and D, which refer to the high-frequency components in the horizontal, vertical, and diagonal directions, respectively. For both the ground-truth image and predicted image, $\mathcal{L}_{DWT}$ compute the MSE between each high-frequency sub-band:

$$\mathcal{L}_{DWT} = \sum_{i=1}^{L} \mathbb{E}[\left\| \hat{H}_i - H_i \right\|^2 + \left\| \hat{V}_i - V_i \right\|^2 + \left\| \hat{G}_i - G_i \right\|^2] \tag{2}$$

where H_i,V_i,D_i are the sub-bands of the ground-truth image in the i-th downsampling, and $\hat{H}_i$,$\hat{V}_i$,$\hat{D}_i$ are the sub-bands of the predicted image in the i-th downsampling, L is the total level of downsampling.

We also add the spatial domain loss named $\mathcal{L}_{GT}$: let the ground-truth image be Y, the predicted image be $\hat{Y}$, and $\mathcal{L}_{GT}$ is the MSE between them:

$$\mathcal{L}_{GT} = \mathbb{E}[\left\| Y - \hat{Y} \right\|^2] \tag{3}$$

The total loss function of pre-trained CNN thus is:

$$\mathcal{L}_{CNN} = \mathcal{L}_{GT} + \alpha\mathcal{L}_{FFT} + \beta\mathcal{L}_{DWT} \tag{4}$$

where α and β are adjustable hyperparameters.

Furthermore, we design a simple CNN using residual-connection (He et al. 2016) and pixel-shuffle (Shi et al. 2016), named **SimpleSR**, for initial prediction (the specific structure is given in the supplementary material). Ablation

studies on the proposed loss function and SimpleSR are given in the supplementary material.

FD-guided Diffusion

After obtaining the image I predicted by the pre-trained CNN, we adapt a diffusion model to predict the residuals between I and the ground truth, i.e., the high-frequency components of the ground-truth image. To this end, we propose a Frequency-Domain guided diffusion (FD-guided diffusion), as shown in Fig.4. In contrast to SR3 (Saharia et al. 2022c), which simply concatenates the bilinear interpolated image with the noisy image x_t at step t, we propose a Frequency-Domain Information Splitter module (FD-Info-Splitter): I and x_t is first fed into the FD-Info-Splitter, whose output is then fed into the U-net (Ronneberger, Fischer, and Brox 2015). We follow the Imagen (Saharia et al. 2022a), where the self-attention layer is added. In addition, a Frequency-Domain guided Cross-Attention mechanism (FD-guild CA) is designed, which utilizes the high-frequency features obtained from DWT at each layer to generate more fine-grained detail features.

FD Info Splitter

For CNN's initial prediction, low-frequency components are mixed with high-frequency contents. As the diffusion model only needs to recover high-frequency details, the input low and high-frequency features have different statuses: the former mainly assist the generation of high-frequency components globally, while the latter is required to provide guidance for fine-grained details in each region. Therefore, we introduce Frequency-Domain Information Splitter (FD Info Splitter), which explicitly separates high-frequency and low-frequency information for better restoration. Additionally, it effectively mitigates noise for noisy images with large time steps, resulting in better noise prediction (The detailed structure of FD Info Splitter is shown in Fig.4).

For the CNN predicted images $x_{cnn} \in \mathbb{R}^{H\times W\times C}$, we first perform 2D FFT along the spatial dimensions to obtain the frequency domain feature map M:

$$M = FFT(x_{cnn}) \in \mathbb{C}^{H\times W\times C} \tag{5}$$

where $FFT(\cdot)$ denotes the 2D FFT. We adapt the methods proposed by (Hu, Shen, and Sun 2018; He et al. 2016) and merged them into the ResSE module (Residual Squeeze-and-Excitation module), the details of which are shown in the supplementary material.

To implement adaptive high-pass filtering, a Gaussian high-pass filter is utilized whose Standard deviation is obtained from M as follows:

$$\sigma = min(|ResSE(M)| + \frac{l}{2}, l) \tag{6}$$

where $l = min(H, W)$. The operation for the acquired $ResSE(M)$ is for numerical stability. After obtaining σ, adaptive gaussian high-pass filter can be given directly as:

$$H(u, v) = 1 - e^{-D^2(u,v)/(2\sigma^2)} \tag{7}$$

where $D(u, v)$ is the distance from the point (u, v) in the frequency domain to the center point. The gaussian high-pass filter are then preformed element-wise multiplication with M to obtain the adaptive high-pass filtered feature map $M^{'}$:

$$M^{'} = A_{hp} \otimes M \tag{8}$$

Finally, we reverse $M^{'}$ back to the spatial domain by adopting inverse FFT to obtain an feature map x_{HF} rich in high-frequency components:

$$x_{HF} = FFT^{-1}(M^{'}) \in \mathbb{R}^{H\times W\times C} \tag{9}$$

where $FFT^{-1}(\cdot)$ denotes the Inverse 2D FFT. Meanwhile, we feed $M^{'}$ into a ResSE module to acquire the attention weights learned in the frequency domain and then perform element-wise multiplication with x_{cnn} to obtain a feature map x_{LF} containing abundant low-frequency information:

$$x_{LF} = ResSE(M) \otimes x_{cnn} \tag{10}$$

These two feature maps, dominated by high-frequency and low-frequency components, are concatenated in the channel dimension. By explicitly separating the input's mixed high-frequency and low-frequency components, the network can utilize both differently and more efficiently.

For a noisy image x_t at a large time step t, the noise components can be so large that it hinders network inference. Hence, an adaptive denoising is utilized on x_t to obtain the partially denoised noisy image $x_t^{'}$:

$$x_t^{'} = ResSE(T) \otimes x_t \tag{11}$$

The three feature maps x_{HF}, x_{LF}, $x_t^{'}$, along with x_{cnn} and x_t, are all concatenated in the channel dimension and fed into the U-net.

HF-guided CA

In the original U-net architecture, the encoder features are directly concatenated with the features obtained by the decoder (Ronneberger, Fischer, and Brox 2015). This fusion facilitates the network to integrate the higher and lower-layer features effectively but lacks the ability to extract high-frequency features. To tackle this issue, we introduce a High-Frequency feature guided Cross-Attention mechanism (HF-guided CA) to recover fine-grained high-frequency details. The flow of the HF-guided CA is illustrated in Fig.4.

We utilize the pre-trained CNN prediction by extracting the $\hat{H}_i$, $\hat{V}_i$, and $\hat{D}_i$ coefficients at the i-th level of the DWT. By adding these extracted coefficients with a linear projection, we obtain the feature map Q with aggregated high-frequency information:

$$Q = Conv_{1\times1}(\hat{H}_i + \hat{V}_i + \hat{D}_i) \tag{12}$$

Then, different linear projections of the input feature map M are constructed to obtain K and V in the cross-attention mechanism (Hou et al. 2019) :

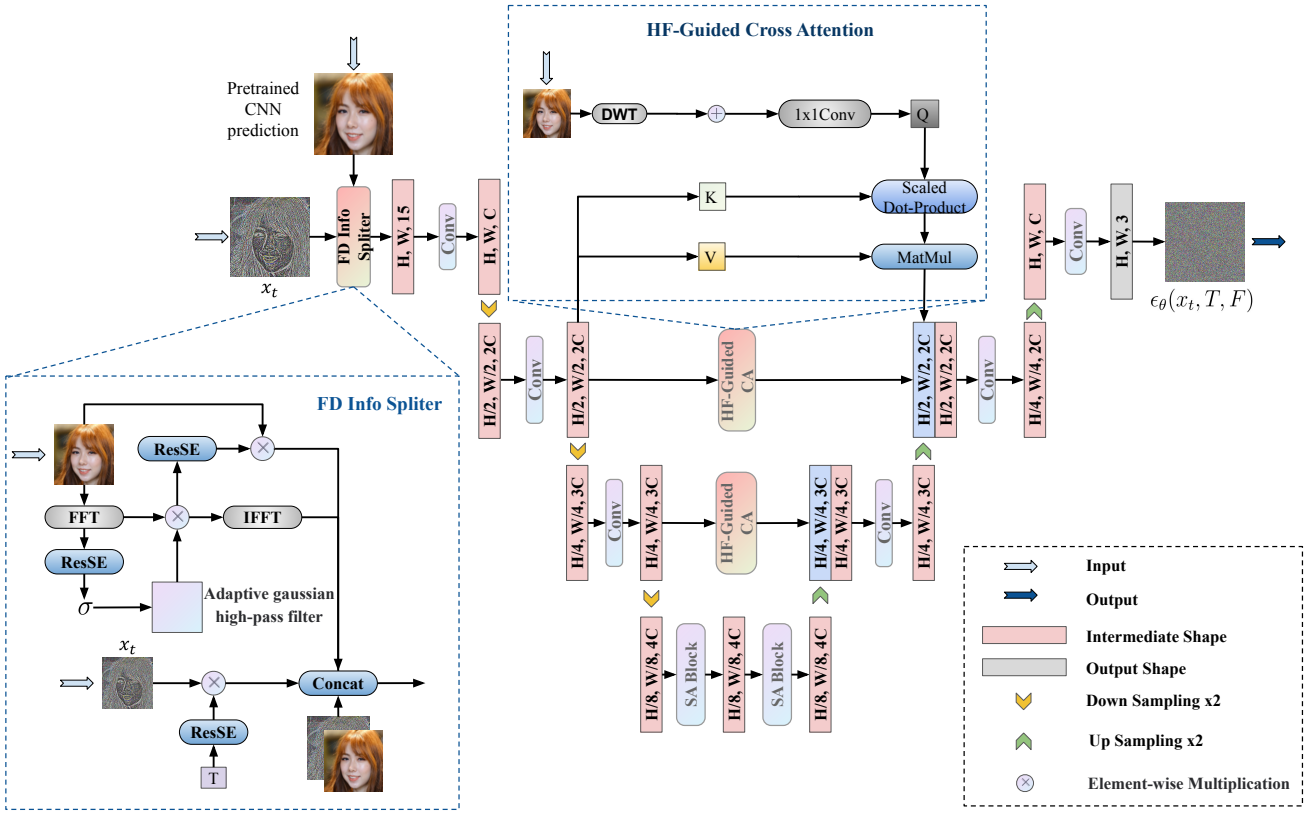

Figure 4: An overview of the model architecture in proposed FD-guided diffusion. The pre-trained CNN prediction and the noisy image x_t from step t are fed into the FD-info-Splitter, and its output is then passed on to a U-net, which is equipped with HF-guided cross-attention.

$$K = Conv_{1\times1}(M) \tag{13}$$

$$V = Conv_{1\times1}(M) \tag{14}$$

The output feature map $M^{'}$ can then be obtained from the formula:

$$M^{'} = Softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{15}$$

where d_k is the number of columns of matrix Q.

Experiments

Performance

To evaluate the performance of our ResDiff model, we compared it with previous diffusion-based and GAN-based methods using four datasets: two face datasets (FFHQ (Karras, Laine, and Aila 2019), and CelebA (Liu et al. 2015)) and two general datasets (Div2k (Agustsson and Timofte 2017), and Urban100 (Huang, Singh, and Ahuja 2015)). The selected evaluation metrics include two distortion-based metrics (PSNR and SSIM (Wang et al. 2004)), as well as a perceptual-based metric (FID (Heusel et al. 2017)). Our ResDiff is trained solely on the provided training data to guarantee a fair comparison. The supplementary material contains detailed information about the training process, hyperparameters, and other relevant details. Since several methods did not state their performance on some datasets we use, their values are marked as "-" in the table. More experiments with different types of datasets are presented in the supplementary material.

FFHQ and CelebA Results The quantitative results at $32\times32 \rightarrow 128\times128$ (4×) ,$256\times256 \rightarrow 1024\times1024$ (4×) on FFHQ (Karras, Laine, and Aila 2019) and $20 \times 20 \rightarrow 160\times160$ (8×), $64\times64 \rightarrow 256\times256$ (4×) on CelebA (Liu et al. 2015) are shown in table 1,2. Our ResDiff demonstrates superior performance compared to all diffusion-based methods, as evidenced by the metrics presented in the table, and has about 50% reduction in Perceptual metrics (FID) than the GAN-based model.

DIV2K and Urban100 Results The quantitative results at $40 \times 40 \rightarrow 160 \times 160$ (4×) on DIV2K (Agustsson and Timofte 2017) and $40 \times 40 \rightarrow 160 \times 160$ (4×) on Urban100 (Huang, Singh, and Ahuja 2015) are shown in table 3. Note that ResDiff's distortion-based metric values can

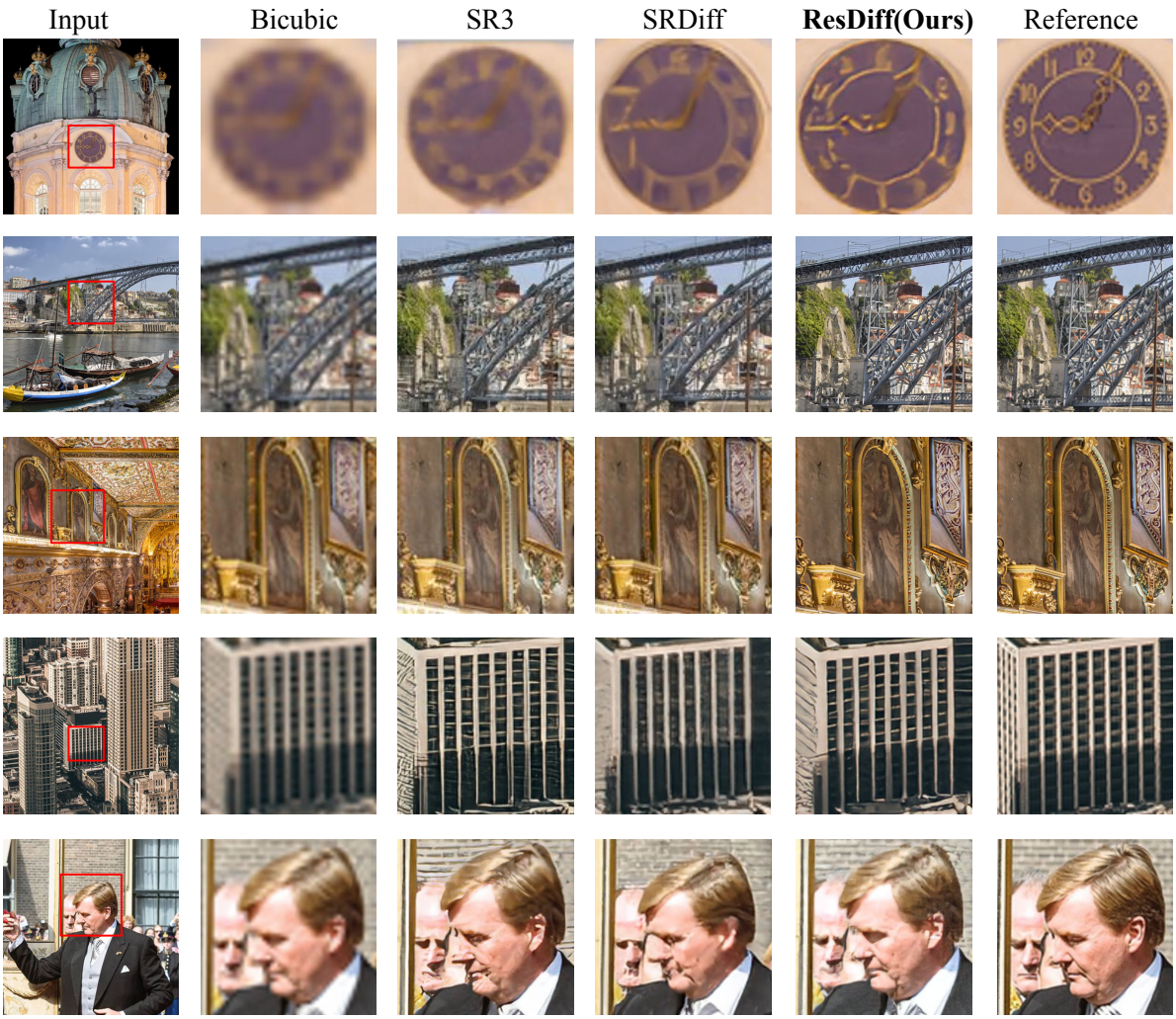

Figure 5: DIV2k 4× results. Note that ResDiff provides richer details and more natural textures than other diffusion-based methods for the recovery of small objects (e.g., the clock in the first column) and difficult scenes (e.g., the bridge structure in the second column, the building in the fourth column).

	32 → 128			256 → 1024		
	PSNR↑	SSIM↑	FID↓	PSNR↑	SSIM↑	FID↓
Ground Truth	∞	1.000	0.00	∞	1.000	0.00
SRGAN	17.57	0.688	156.07	21.49	0.515	60.67
ESRGAN	15.43	0.267	166.36	19.84	0.353	72.73
BRGM	24.16	0.70	-	-	–	-
PULSE	15.74	0.37	-	-	-	-
SRDiff	26.07	0.794	72.36	23.01	0.656	56.17
SR3	25.37	0.778	75.29	22.78	0.647	60.12
ResDiff	**26.73**	**0.818**	**70.54**	**23.15**	**0.668**	**53.23**

Table 1: Quantitative comparison on the FFHQ (Karras, Laine, and Aila 2019) dataset, where the bolded values represent the best value in each evaluation metric.

	20 → 160			64 → 256		
	PSNR↑	SSIM↑	FID↓	PSNR↑	SSIM↑	FID↓
Ground Truth	∞	1.000	0.00	∞	1.000	0.00
ESRGAN	23.24	0.66	-	-	-	-
PULSE	-	-	-	22.74	0.623	40.33
SRFlow	25.28	0.72	-	-	-	-
SRDiff	25.32	0.73	80.98	26.84	0.792	39.16
SR3	24.89	0.728	83.11	26.04	0.779	43.27
ResDiff	**25.37**	**0.734**	**78.52**	**27.16**	**0.797**	**38.47**

Table 2: Quantitative comparison on the CelebA (Liu et al. 2015) dataset, where the bolded values represent the best value in each evaluation metric.

	DIV2K 4×			Urban100 4×		
	PSNR↑	SSIM↑	FID↓	PSNR↑	SSIM↑	FID↓
Ground Truth	∞	1.000	0.00	∞	1.000	0.00
SRDiff	26.87	0.69	110.32	26.49	0.79	51.37
SR3	26.17	0.65	111.45	25.18	0.62	61.14
ResDiff	**27.94**	**0.72**	**106.71**	**27.43**	**0.82**	**42.35**

Table 3: Quantitative comparison on the DIV2K (Agustsson and Timofte 2017) and Urban100 (Huang, Singh, and Ahuja 2015) dataset, where the bolded values represent the best value in each evaluation metric.

significantly outperform other diffusion-based methods on these general datasets whose restoration is more difficult. Fig.5 presents partial results of ResDiff and other diffusion-based methods.

Ablation Study

In this section, we perform an ablation study on FFHQ (4×) to investigate the effectiveness of each component in ResDiff, including the influence of different CNNs, and the usefulness of the proposed FD Info Splitter/HF-guided CA. The results are shown in Table 4. Note that utilizing the residual structure, even with a simple bilinear interpolation for the initial prediction, can significantly improve the performance. In terms of CNN selection, our proposed SimpleSR also outperforms SRCNN (Dong et al. 2014). Moreover, the addition of FD Info Splitter and HF-guided CA both have an improvement in the results. More detailed ablation studies are given in the supplementary material.

Model Components			Metrics		
CNN	FD Info Splitter	HF-guided CA	PSNR↑	SSIM↑	FID↓
SimpleSR	✓	✓	26.73	0.818	70.54
N/A	✓	✓	25.49	0.781	74.18
Bilinear	✓	✓	25.99	0.792	74.29
SRCNN	✓	✓	26.14	0.809	72.17
SimpleSR (only $\mathcal{L}_{GT}$)	✓	✓	26.47	0.812	71.58
SimpleSR			25.41	0.788	77.21
SimpleSR		✓	26.09	0.796	72.42
SimpleSR	✓		25.97	0.793	73.17

Table 4: Ablation study over different model components on the ffhq (Karras, Laine, and Aila 2019) test sets (The model components we use are placed in the first row). N/A denotes no residual structure used.

Conclusion and Future Work

In this paper, we propose ResDiff, a residual structure-based diffusion model. In contrast to the previous works, which only adapt LR images to generate HR images, ResDiff utilizes the feature-richer CNN prediction for guidance. Meanwhile, we introduce a frequency-domain-based loss function to the CNN and design a frequency-domain guided diffusion to facilitate the diffusion model in generating low-frequency information. Comprehensive experiments on different datasets demonstrate that the proposed ResDiff accelerates the training convergence speed and provides superior image generation quality.

Our ResDiff can also be adapted for other image restoration tasks, such as image blind super-resolution, deblurring, and inpainting. Although ResDiff can accelerate convergence, operations such as DWT are still time-consuming and call for optimization in future work. In addition, it can be seen from the supplementary material that the color will appear a large discrepancy when the model is under-trained, which may be caused by a lack of color features in the guided high-frequency information. Utilizing a global color feature may well address this issue in future work. Moreover, our ResDiff does not outperform current State-Of-The-Art(SOTA) SISR methods (Chen et al. 2022; Zhang et al. 2022). This is attributed to the disparity between model parameters. Due to equipment limitations, adopting a larger U-net model in ResDiff is left to future work. In addition, if a pre-trained SOTA model is applied to replace the CNN in ResDiff, it may be possible to establish a new SOTA. Finally, ResDiff may consider incorporating more DPM techniques (Rombach et al. 2022; Dhariwal and Nichol 2021; Ho and Salimans 2022) and superior network architectures (Peebles and Xie 2022; Chen et al. 2021) in the future.

Acknowledgments

We gratefully thank the creators of the dataset and the server support from Shandong University and Linyi University. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFB4500602, the Key Research and Development Program of Jiangsu Province under Grant BE2021093, Distinguished Young Scholar of Shandong Province under Grant ZR2023JQ025, Taishan Scholars Program under Grant tsqn202211290, and Major Basic Research Projects of Shandong Province under Grant ZR2022ZD32.

References

Agustsson, E.; and Timofte, R. 2017. NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*, 1122–1131. IEEE Computer Society.

Bai, C.; Zhang, M.; Zhang, J.; Zheng, J.; and Chen, S. 2022. LSCIDMR: Large-Scale Satellite Cloud Image Database for Meteorological Research. *IEEE Transactions on Cybernetics*, 52(11): 12538–12550.

Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A. L.; and Zhou, Y. 2021. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *CoRR*, abs/2102.04306.

Chen, X.; Wang, X.; Zhou, J.; and Dong, C. 2022. Activating More Pixels in Image Super-Resolution Transformer. *CoRR*, abs/2205.04437.

Choi, J.; Kim, S.; Jeong, Y.; Gwon, Y.; and Yoon, S. 2021. ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 14347–14356. IEEE.

Cooley, J. W.; and Tukey, J. W. 1965. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19: 297–301.

Deng, X.; Yang, R.; Xu, M.; and Dragotti, P. L. 2019. Wavelet Domain Style Transfer for an Effective Perception-Distortion Tradeoff in Single Image Super-Resolution. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 3076–3085. IEEE.

Dhariwal, P.; and Nichol, A. Q. 2021. Diffusion Models Beat GANs on Image Synthesis. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 8780–8794.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2014. Learning a Deep Convolutional Network for Image Super-Resolution. In Fleet, D. J.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV*, volume 8692 of *Lecture Notes in Computer Science*, 184–199. Springer.

Dou, J.; Tu, Z.; and Peng, X. 2020. Single Image Super-resolution Reconstruction with Wavelet based Deep Residual Learning. In *2020 Chinese Control And Decision Conference (CCDC)*, 4270–4275.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; and Bengio, Y. 2014. Generative Adversarial Nets. In *NIPS*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 770–778. IEEE Computer Society.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 6626–6637.

Ho, J.; and Salimans, T. 2022. Classifier-Free Diffusion Guidance. *CoRR*, abs/2207.12598.

Hou, R.; Chang, H.; Ma, B.; Shan, S.; and Chen, X. 2019. Cross Attention Network for Few-shot Classification. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 4005–4016.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-Excitation Networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 7132–7141. Computer Vision Foundation / IEEE Computer Society.

Huang, J.; Singh, A.; and Ahuja, N. 2015. Single image super-resolution from transformed self-exemplars. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 5197–5206. IEEE Computer Society.

Karras, T.; Laine, S.; and Aila, T. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 4401–4410. Computer Vision Foundation / IEEE.

Kawar, B.; Elad, M.; Ermon, S.; and Song, J. 2022. Denoising Diffusion Restoration Models. In *ICLR Workshop on Deep Generative Models for Highly Structured Data (ICLRW)*.

Ledig, C.; Theis, L.; Huszar, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A. P.; Tejani, A.; Totz, J.; Wang, Z.; and Shi, W. 2017. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 105–114. IEEE Computer Society.

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.

Liang, J.; Lugmayr, A.; Zhang, K.; Danelljan, M.; Gool, L. V.; and Timofte, R. 2021. Hierarchical Conditional Flow: A Unified Framework for Image Super-Resolution and Image Rescaling. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 4056–4065. IEEE.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 3730–3738. IEEE Computer Society.

Lugmayr, A.; Danelljan, M.; Gool, L. V.; and Timofte, R. 2020. SRFlow: Learning the Super-Resolution Space with Normalizing Flow. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part V*, volume 12350 of *Lecture Notes in Computer Science*, 715–732. Springer.

Mallat, S.; and Hwang, W. 1992. Singularity detection and processing with wavelets. *IEEE Transactions on Information Theory*, 38(2): 617–643.

Mirchandani, K.; and Chordiya, K. 2021. DPSRGAN: Dilation Patch Super-Resolution Generative Adversarial Networks. In *2021 6th International Conference for Convergence in Technology (I2CT)*, 1–7.

Peebles, W.; and Xie, S. 2022. Scalable Diffusion Models with Transformers. *CoRR*, abs/2212.09748.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *ArXiv*, abs/2204.06125.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 10674–10685. IEEE.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Navab, N.; Hornegger, J.; III, W. M. W.; and Frangi, A. F., eds., *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, volume 9351 of *Lecture Notes in Computer Science*, 234–241. Springer.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S. K. S.; Ayan, B. K.; Mahdavi, S. S.; Lopes, R. G.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022a. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *CoRR*, abs/2205.11487.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, S. K. S.; Ayan, B. K.; Mahdavi, S. S.; Lopes, R. G.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022b. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *ArXiv*, abs/2205.11487.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022c. Image Super-Resolution Via Iterative Refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–14.

Shi, W.; Caballero, J.; Huszar, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 2016. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 1874–1883. IEEE Computer Society.

Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2018a. Recovering Realistic Texture in Image Super-Resolution by Deep Spatial Feature Transform. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 606–615. Computer Vision Foundation / IEEE Computer Society.

Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; and Loy, C. C. 2018b. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In Leal-Taixé, L.; and Roth, S., eds., *Computer Vision - ECCV 2018 Workshops - Munich, Germany, September 8-14, 2018, Proceedings, Part V*, volume 11133 of *Lecture Notes in Computer Science*, 63–79. Springer.

Wang, Y.; Yu, J.; and Zhang, J. 2023. Zero-Shot Image Restoration Using Denoising Diffusion Null-Space Model. In *The Eleventh International Conference on Learning Representations(ICLR)*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4): 600–612.

Whang, J.; Delbracio, M.; Talebi, H.; Saharia, C.; Dimakis, A. G.; and Milanfar, P. 2022. Deblurring via Stochastic Refinement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 16272–16282. IEEE.

Zhang, D.; Huang, F.; Liu, S.; Wang, X.; and Jin, Z. 2022. SwinFIR: Revisiting the SwinIR with Fast Fourier Convolution and Improved Training for Image Super-Resolution. *CoRR*, abs/2208.11247.

Zhang, W.; Liu, Y.; Dong, C.; and Qiao, Y. 2019. RankSRGAN: Generative Adversarial Networks With Ranker for Image Super-Resolution. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 3096–3105. IEEE.