This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Diffusion Transformer meets Multi-level Wavelet Spectrum for Single Image Super-Resolution

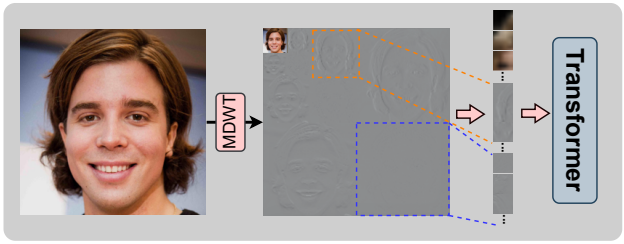
Peng Du[1], Hui Li[1]*, Han Xu[1], Paul Barom Jeon[2], Dongwook Lee[2], Daehyun Ji[2], Ran Yang[1], Feng Zhu[1]
[1]Samsung R&D Institute China Xi'an (SRCX)
[2]Samsung Electronics Co., LTD., South Korea
peng03.du, hui01.li, han.xu, paul.barom.jeon, dw23.lee, derek.ji, ran01.yang, f15.zhu@samsung.com

Abstract

Discrete Wavelet Transform (DWT) has been widely explored to enhance the performance of image super-resolution (SR). Despite some DWT-based methods improving SR by capturing fine-grained frequency signals, most existing approaches neglect the interrelations among multi-scale frequency sub-bands, resulting in inconsistencies and unnatural artifacts in the reconstructed images. To address this challenge, we propose a Diffusion Transformer model based on image Wavelet spectra for SR (DTWSR). DTWSR incorporates the superiority of diffusion models and transformers to capture the interrelations among multi-scale frequency sub-bands, leading to a more consistence and realistic SR image. Specifically, we use a Multi-level Discrete Wavelet Transform (MDWT) to decompose images into wavelet spectra. A pyramid tokenization method is proposed which embeds the spectra into a sequence of tokens for transformer model, facilitating to capture features from both spatial and frequency domain. A dual-decoder is designed elaborately to handle the distinct variances in low-frequency (LF) and high-frequency (HF) sub-bands, without omitting their alignment in image generation. Extensive experiments on multiple benchmark datasets demonstrate the effectiveness of our method, with high performance on both perception quality and fidelity.

1. Introduction

Single-Image Super-Resolution (SISR) has gained growing attention for decades because of its broad application. It restores high-resolution (HR) images from the given low-resolution (LR) inputs, aiming at high performance on both objective fidelity and perceptual quality. Most methods establish the mapping from LR to HR images in image pixel domain. To capture the fine-grained frequency details critical for SR, some approaches use the Discrete Wavelet Transform (DWT) to convert images into the frequency domain. DWT depicts an image by a series of frequency sub-bands. The low-frequency (LF) sub-band reflects image global topology and affects objective fidelity, while the high-frequency (HF) sub-bands represent image textural details and affect perceptual quality significantly [11, 30]. As indicated in [21], SISR can be formulated as a wavelet coefficients prediction task. With correct predicting of wavelet coefficients based on LR input, the HR image can be reconstructed via inverse DWT (IDWT). Explicit optimization of wavelet coefficients in the frequency domain shows



(a) The pixel image is represented by multi-level wavelet spectra. Transform is explored to model the complex relations among the multi-scale frequencies.



(b) SR result w/o multi-scale frequency interrelations

(c) Ours (with multi-scale frequency interrelations considered)

(d) Ground truth

Figure 1. A transformer model based on multi-level wavelet spectra is explored for SR, enabling the learning of multi-scale frequency relationships to enhance SR results. Comparing (b) and (c), the proposed method produces more natural textures.

*Corresponding author.

enhanced generation quality [21, 30].

Plenty of works were proposed to improve the prediction accuracy of wavelet coefficients. Wavelet-SRNet [21] adopts N independent CNN subnets to predict multi-level of wavelet coefficients in parallel. WaveFace [41] employs a U-Net based model to recover the HF sub-bands sequentially during the upsampling process. WFEN [33] takes DWT on each feature level in U-Net model to mitigate feature distortion during downsampling process. Nevertheless, to the best of our knowledge, existing methods generally treat each level of HF coefficients independently, without considering the interrelations among the multi-scale HF sub-bands.

Taking the objective of SISR into account, not only the spatial topology is important to avoid spatial distortion, the correlation among image frequency sub-bands is also crucial for better perceptual quality. Hence in this work, we explore a transformer model grounded on multi-level wavelet spectra for SISR, leveraging the excellence of transformer in modeling complex long-range relationships. Our model enables to uncover the correlations among the multiple scale frequency sub-bands, leading to elaborate textural details (as illustrated in Fig. 1 (b) *vs*. (c)).

Specially, we adopt Mallat decomposition [40] for multi-level Discrete Wavelet Transform (MDWT), which decomposes the LF sub-band repeatedly at each subsequent level (detailed in Sec. 3). The obtained frequency sub-bands are then combined together as a wavelet spectra representation of the image. As shown in Fig. 1, it consists of one LF sub-band and multiple HF sub-bands at different scales, containing distinct levels of textural information. We then split the wavelet representation into patches for token embedding, not only on LF sub-band, but also on HF sub-bands. Unlike conventional methods that partition images spatially, our method divides images from the viewpoint of both spatial and frequency domains, facilitating the learning the frequency relationships among sub-bands. In addition, we propose a pyramid tokenization method given the sparsity of HF sub-bands. It reduces the token numbers largely and saves computation in transformer calculation without compromising model performance.

Inspired by the outstanding capability of Diffusion Model (DM) in generating fine image details, we formulate our method by using the conditional diffusion framework, and propose a **D**iffusion **T**ransformer model based on image **W**avelet spectra for SISR, abbreviated as **DTWSR**. DMs reverse a diffusion process iteratively to achieve high-quality mapping from randomly sampled Gaussian noise to target images, avoiding the instability and mode-collapse present in previous generative models [20, 45, 53]. Due to the distinct variances in MDWT sub-bands, particularly between the smooth LF and sparse HFs, it is challenging to use a unified transformer model to denoise both LF and HF sub-bands simultaneously. Therefore, we design a dual-decoder transformer model, one for generating the high-energy elementary contents in LF (named as LEDec) and the other one for generating the sparse HF details (named as HDDec). It should be noted that the elementary contents from LEDec is not equal to LF sub-band. The LF sub-band still has HF components, though quite few. (A simple understanding is that we can continue the wavelet transform on LF to peel off the included HF components.) Thus, HDDec is designed to produce both the multi-level HF sub-bands and the HF components of LF sub-band. On one hand, our design is able to capture the interrelations among multi-scale HF sub-bands. On other hand, it promotes the realignment between LF and HF sub-bands, achieving SR with improved fidelity and perceptual quality.

The main contributions of this paper are as follows:

- We propose a diffusion transformer model based on image wavelet spectra for SISR. It enables to explore the correlations among multi-scale frequency sub-bands.
- We design a pyramid tokenization method for embedding the multi-scale wavelet spectra. It reduces the token number largely for efficient calculation.
- A dual-decoder model is designed to prevent the entanglement of smooth and sparse frequency distributions for better fidelity and finer details.
- Extensive experiments are conducted on key benchmarks for face and general image SR tasks. Our method exhibits state of-the-art qualitative and quantitative results with improved image fidelity and perceptual quality.

2. Related work

SISR has achieved great progress with the development of deep learning, including both model architecture and training framework. To improve the visual quality, various generative models are applied to train SISR model, including GAN [5, 7, 16, 49, 50], flow models [36, 38, 63] and Diffusion Models (DM) [22, 41, 43, 51, 55]. Our work applies diffusion transformer based on wavelet spectra for SISR.

Diffusion based SISR. DM is rising as a powerful solution for high-quality image generation. SR3 [46] adapts conditional DMs by concatenating upsampled LR with noisy HR image to perform SISR task. To speed up convergence and stabilize the training of DMs, SRDiff [31] introduces residual prediction to speed up the convergence of DMs. ResDiff [47] uses a CNN network for initial recovery and then refines textural details by DM. IDM [15] and ASIG [28] explore DMs in continuous SISR by integrating implicit neural representation. ResShift [56] and SinSR [51] accelerate the inference speed of DM by modifying its sampling process and using knowledge distillation, respectively.

Discrete Wavelet Transform (DWT) based SISR. DWT has been used widely in SISR given its ability to express frequency information [11, 24, 30, 32, 43, 54]. DWSR [17]

and DiWa [43] are built on single-level DWT to improve model on precise textural details. Wavelet-SRNet [21] and JWN [64] uses multi-branch CNN layers to predict wavelet coefficients based on LR input. WaveMixSR [23], WTRN [34] and WFEN [33] apply DWT on the extracted image features to complement HF information in SISR. WaveFace [41] and WaveDM [22] leverages the exponential shrinking of image size after DWT to reduce the computation burden of DM. Deng *et al*. [11] proposed wavelet domain style transfer to achieve better perception-distortion (PD) trade-off for SISR. PDASR [62] and WGSR [30] optimize the loss on wavelet sub-bands to improve PD trade-off.

Transformer based SISR. Transformer based models are explored in SISR given its long-range modeling ability. SwinIR [35] applies Swin transformer for image restoration. SwinFIR [59] improves SwinIR by incorporating Fourier Convolution to capture global information. HAT [6] combines self-attention, channel attention and overlapping cross-attention to active more pixel for better SR. Restormer [57] proposes to perform self-attention in channel direction to capture long-range pixel interactions and achieves high performance in image restoration. LMLT [29] divides image features along the channel dimension and employs attention with varying feature sizes to capture both local and global information. These works are based on pixel domain images. DWT is often employed in attention blocks of transformer model to enhance image feature, like [2, 13, 33, 37]. To our knowledge, we are the first to model image's multi-scale wavelet spectra by using the basic transformer architecture for SR task.

3. Discrete wavelet transform

The discrete wavelet transform (DWT) is widely used to decompose an image into LF and HF sub-bands, especially the Haar wavelet [18] used in this paper.

Given a pixel image $I \in \mathbb{R}^{H\times W\times 3}$, we decompose it by DWT operation ($\mathrm{DWT}(\cdot)$), and thus the low-frequency sub-band $x_L \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times 3}$ and high-frequency sub-bands $\{x_V, x_H, x_D\} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times 3}$ can be produced:

$$x_L^1, x_V^1, x_H^1, x_D^1 = \mathrm{DWT}(I). \tag{1}$$

The process can be conducted once more on x_L^1, resulting in

$$x_L^2, x_V^2, x_H^2, x_D^2 = \mathrm{DWT}(x_L^1). \tag{2}$$

By continuing the process, we have $\{x_L^J, x_V^J, x_H^J, x_D^J\} \in \mathbb{R}^{\frac{H}{2^J}\times\frac{W}{2^J}\times 3}$ after the J-th DWT.

Replacing the LF sub-band recursively by the decomposed sub-bands in each level [40], the outputs after J-th DWT are $\{x_L^J, x_V^J, x_H^J, x_D^J, x_V^{J-1}, ..., x_D^1\}$. We reshape the multi-level sub-bands together and form a J-level wavelet spectrum representation of the image, denoted as I_J^{fre}, *i.e.*,

$$I_J^{fre} = \mathrm{MDWT}(I, J). \tag{3}$$

Reversibly, the pixel image I can be reconstructed via J-th invert DWT, (denoted as IMDWT):

$$I = \mathrm{IMDWT}(I_J^{fre}, J). \tag{4}$$

An example is presented in Fig. 1a.

4. Methodology

In this section, we will introduce our Diffusion Transformer model based on Wavelet spectra for SR task.

4.1. Conditional DM on wavelet spectra for SISR

DM is a parameterized Markov chain that produces samples matching the training data distribution. It consists of a forward diffusion process and a reverse denoising process. The diffusion process gradually adds Gaussian noise to a clean image according to a pre-defined Markov process, while the denoising process recovers the clean image from Gaussian noise by removing noise iteratively via a denoising network learned from the diffusion process. For SR task, it is request that the recovered image is consistent with the content from the given LR input, resulting in a conditional DM:

$$p_\theta(I_{t-1}|I_t, I_{lr}) = \mathcal{N}(I_{t-1}; \mu_\theta(I_t, t, I_{lr}), \mathbf{\Sigma}_\theta(I_t, t, I_{lr})), \tag{5}$$

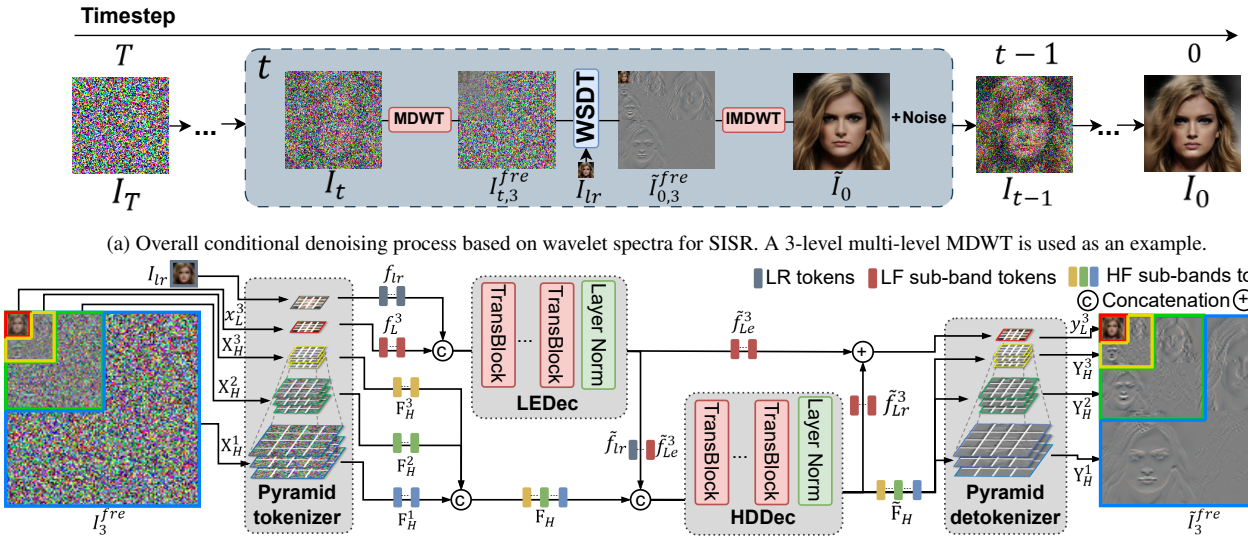
where I_{lr} denotes the LR input, θ is the parameter of our designed Wavelet Spectrum Denoising network with Transformer, named as **WSDT**, and t is the denoising step. By refining I_t recursively conditioning on I_{lr}, the SR image I_0 can be obtained. The process is illustrated in Fig. 2a.

Different from previous methods that remove noise from I_t in pixel domain, we attempt to leverage the wavelet frequency spectrum to improve the generation quality. Hence we transform the pixel image I_t to a J-level wavelet spectrum representation $I_{t,J}^{fre}$ for refinement (as shown in Fig. 2a). The included LF sub-band and the set of HF sub-bands are denoted as $x_{t,L}^J$ and $\mathbf{X}_{t,H} = \{\mathbf{X}_{t,H}^j\}$ separately, where $\mathbf{X}_{t,H}^j = \{x_{t,V}^j, x_{t,H}^j, x_{t,D}^j\}, j \in \{1, ..., J\}$.

In our model, we set the level of MDWT according to the magnification of SR. For a upscaling factor N, we perform J-level DWT, with $J = \mathrm{ceil}(\log_2 N)$. With this setting, the size of LF sub-band will be no larger than that of LR input. It would be relatively easier to learn the mapping between two similar size images with similar distribution.

4.2. Wavelet Spectrum Denoising Network with Transformer (WSDT)

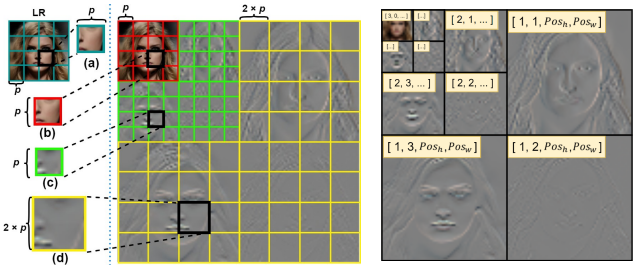
Fig. 2b presents the architecture of WSDT. Given the noisy image in wavelet spectrum $I_{t,J}^{fre}$, we firstly patchify it into a sequence of tokens. Then dual transformer decoders are designed to denoise the elementary contents in LF and the multi-scale HF details respectively, with in-context conditioning on LR input. The transformer blocks will learn the interrelations both in spatial domain and among multi-scale



(a) Overall conditional denoising process based on wavelet spectra for SISR. A 3-level multi-level MDWT is used as an example.

(b) Detailed illustration of WSDT in (a), where the timestep condition and details of decoders are simplified for conciseness.

Figure 2. Overview of the DTWSR framework. (a) shows SR sampling process, which follows the classic conditional denoising process of diffusion model (DM). The gray box shows how conditional DM on wavelet spectra is applied in each step. (b) illustrates the detailed structure of the proposed denoising network WSDT. The spectra image is embedded by a pyramid tokenizer. LEDec denoises the LF sub-band to obtain $\tilde{f}^3_{Le}$ under the guidance of the LR image. HDDec decodes the HF sub-bands tokens $\tilde{\mathbf{F}}_H$ and refines $\tilde{f}^3_{Le}$ by adding LR residual $\tilde{f}^3_{Lr}$, conditioning on LR features. Finally, the pyramid detokenizer transforms LF and HF tokens into the denoised spectrum $\tilde{I}^{fre}_3$.



(a) Pyramid patch sizes (b) 4D positional encoding

Figure 3. Illustration of pyramid tokenization. Our method enables consistent receptive fields across frequency sub-bands.

frequencies, leading to more accurate denoising across sub-bands. In this section, we omit the time step t for simplicity as the operations are the same for each time step.

4.2.1. Pyramid tokenization

Conventional methods split image into same size patches for embedding [12]. Considering the sparsity of HF components in wavelet spectrum image I^{fre}_J, we design a pyramid tokenization method. The LF sub-band is divided using a smaller patch size, while the sparse HF sub-bands are split by a larger patch size, as shown in Fig. 3a. Moreover, in order to keep consistent receptive fields across different level of sub-bands, we define the pyramid patch size p^j according to its levels j in MDWT:

$$p^j = p_{min} \times 2^{J-j}, j \in \{1, ..., J\}\,, \tag{6}$$

where p_{min} is the patch size for LF sub-band.

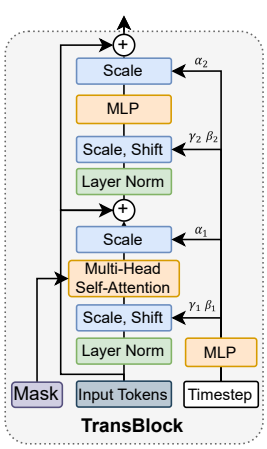
The pyramid tokenization is achieved by convolutional layers $\mathrm{Conv2d}(\cdot)$, with the kernel size and stride set to be p^j. For each level of $\mathbf{X}^j_H$, we concatenate them together for embedding. LR input and LF sub-band are embedded using separate CNN layers. The resulted image tokens are denoted as $\{f_{lr}, f^J_L, \mathbf{F}_H\}, \mathbf{F}_H = \{\mathbf{F}^j_H\}, j \in \{J, ..., 1\}$:

$$\begin{aligned} f_{lr} &= \mathrm{Conv2d}_{lr}(I_{lr}), f^J_L = \mathrm{Conv2d}_L(x^J_L), \\ \mathbf{F}^j_H &= \mathrm{Conv2d}^j_H(\mathbf{X}^j_H), j \in \{J, ..., 1\}\,. \end{aligned} \tag{7}$$

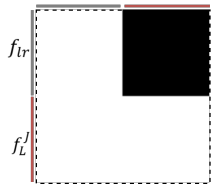
Next, we define the position embedding for each token. With our pyramid tokenization, the resulted tokens have the same 2D-absolute position $[Pos_h, Pos_w]$ in each sub-band, which makes it easier to learn the relations among sub-bands. To distinguish the level of wavelet spectrum and the specific sub-band in each level, we additionally specify the level $j \in \{1, ..., J\}$ and the sub-band position $d \in \{x_L = 0, x_V = 1, x_D = 2, x_H = 3\}$, leading to a 4D position $[j, d, Pos_h, Pos_w]$ for each token, as the example shown in Fig. 3b. The 4D position is encoded by standard ViT frequency-based positional embeddings (the sine-cosine version) [12] and then added to the patch embeddings to retain positional information.

4.2.2. Dual-decoder design

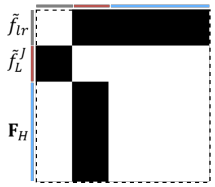
Given that x^J_L and $\{\mathbf{X}^j_H\}$ have different distributions, using a unified decoder to denoise spectra with variant distributions is difficult. Thus, we design dual transformer decoders. Rather than denoising x^J_L and $\{\mathbf{X}^j_H\}$ separately



(a) TransBlock architecture



(b) Self-attention mask M_{low} in LEDec



(c) Self-attention mask M_{high} in HDDec

Figure 4. Illustration of TransBlock and the designed masks in decoders. The black parts correspond to the masked tokens in self-attention computation. Gray lines indicate LR tokens. Red lines indicate LF sub-band tokens. Blue lines indicate HF sub-bands tokens.

as in previous methods [22, 41], we handle their components more carefully. Considering the HF component left in x_L^J, we use one decoder to denoise the smooth elementary contents in x_L^J. The decoder is named as LF Elementary Decoder (**LEDec**). The other one will denoise all the HF coefficients $\{\mathbf{X}_H^j\}$ as well as the left HF components in x_L^J (denoted as LF Residual). The decoder is named as HF Detail Decoder (**HDDec**). Both decoders are composed of multiple transformer blocks [44] as denoted by TransBlock in Fig. 2. The detailed structure of each TransBlock is presented in Fig. 4a.

LEDec. As shown in Fig. 2b, LEDec aims to denoise the smooth elementary component in x_L^J. We use the in-context conditioning method [44] to incorporate the information from LR input. In particular, we concatenate the LR tokens f_{lr} and noised LF tokens f_L^J as input to LEDec. To prevent LR condition from being contaminated by the noised LF tokens, we tailor the attention mask M_{low} in self-attention computation shown in Fig. 4b. The time step t is also embedded and participates in the denoising process via Adaptive layer norm zero (AdaLN-Zero) manner [44] in each TransBlock. The process can be formulated as:

$$\tilde{f}_{lr}, \tilde{f}_{Le}^J = \text{LEDec}([f_{lr}, f_L^J], M_{low}, t), \tag{8}$$

where $[\cdot,\cdot]$ is the concatenation operation.

HDDec. As depicted in Fig. 2b, the inputs to HDDec include the encoded LR and LF tokens, as well as the noised HF tokens $\mathbf{F}_H$. HDDec denoises the multi-level HF sub-bands $\{\mathbf{X}_H^j\}$ and LF Residual. The denoising of LF Residual not only supplements the HF components in x_L^J, but also promotes realignment of LF and HF sub-bands, contributing to finer image generation. We use in-context conditioning manner as well for embedding LR information. A tailored attention mask M_{high} is designed as illustrated in Fig. 4c, which will avoid the unnecessary interaction among tokens. The operations are formulated as

$$\hat{f}_{lr}, \tilde{f}_{Lr}^J, \tilde{\mathbf{F}}_H = \text{HDDec}([\tilde{f}_{lr}, \tilde{f}_{Le}^J, \mathbf{F}_H], M_{high}, t), \tag{9}$$

where M_{high} is worth to be noticed: (1) LR tokens are invisible to LF tokens to avoid LR conditioning in HDDec to be used for LF sub-band denoising and force the realignment between LF and HF sub-bands. (2) LF tokens are invisible to HF tokens to prevent the influence from LF tokens on HF.

Detokenization. Finally, the obtained tokens will be detokenized into wavelet coefficients. For LF sub-band, $\tilde{f}_{Le}^J$ and $\tilde{f}_{Lr}^J$ are added together for detokenizing, while HF coefficients are detokenized from $\tilde{\mathbf{F}}_H = \{\tilde{\mathbf{F}}_H^j\}$ respectively. We apply a layer norm (AdaLN) to incorporate the time step t, and decode each token linearly by $\text{FC}(\cdot)$ into a $p^j \times p^j \times c$ tensor, c is the channel number of that spectrum *i.e.*,

$$\begin{aligned} y_L^J &= \text{FC}_L(\tilde{f}_{Le}^J + \tilde{f}_{Lr}^J, t) \\ \mathbf{Y}_H^j &= \text{FC}_H^j(\mathbf{F}_H^j, t), j \in \{1, ..., J\}\,, \end{aligned} \tag{10}$$

The output tensors are then rearranged according to their original spatial layout, resulting $\tilde{I}_J^{fre} = \{y_L^J, \mathbf{Y}_H\}$, $\mathbf{Y}_H = \{\mathbf{Y}_H^j\}, j \in \{1, ..., J\}$. It is transformed to pixel image by inverse wavelet transform, *i.e.*, $\tilde{I} = \text{IMDWT}(\tilde{I}_J^{fre}, J)$.

4.2.3. Optimization.

To accelerate the denoising process, we adopt the optimization method proposed in DDGAN [53]. It introduces a time-dependent discriminator to learn data distribution at a large denoising step (which is no longer Gaussian), enabling fast sampling without affecting model convergence. Specifically, let I_0 be the clean HR image and I_t be a noised image at timestep t sampled from the diffusion process $q(I_t|I_0)$:

$$q(I_t|I_0) = \mathcal{N}(I_t; \sqrt{\alpha_t} I_0, (1-\alpha_t)\mathbf{I}), \tag{11}$$

where α_t is predefined according to noise schedule.

During the denoising process, our network outputs an denoised image $\tilde{I}_0$ in each step, which is an approximation of I_0. A perturbed sample $\tilde{I}_{t-1}$ can be derived by Eq. (11). DDGAN trains a discriminator $D(\cdot)$ to distinguish the real pairs (I_{t-1}, I_t) and the fake pairs $(\tilde{I}_{t-1}, I_t)$ adversarially, formulated as:

$$\begin{aligned} \mathcal{L}_{adv}^D &= -\log(D(I_{t-1}, I_t, t)) + \log(D(\tilde{I}_{t-1}, I_t, t)). \\ \mathcal{L}_{adv}^G &= -\log(\text{D}(\tilde{I}_{t-1}, I_t, t)). \end{aligned} \tag{12}$$

To preserve the consistency of wavelet sub-bands without losing of frequency details, we build reconstruction term by L_1 loss in both pixel and frequency domain:

$$\mathcal{L}_{pixel} = ||\tilde{I}_0 - I_0||, \quad \mathcal{L}_{fre} = ||\tilde{I}_{0,J}^{fre} - I_{0,J}^{fre}||. \tag{13}$$

The overall objective of the generator is

$$\mathcal{L}^G = \alpha \mathcal{L}_{adv}^G + \beta \mathcal{L}_{pixel} + \gamma \mathcal{L}_{fre}, \tag{14}$$

where α, β, γ are adjustable weighting hyper-parameters.

5. Experiments

5.1. Implementation details

Datasets. We evaluate our method on face and general scene datasets. For face SISR, we train the proposed DTWSR on FFHQ [26] and evaluate on CelebA [25] validation set. For general scene SISR, we use DIV2K [1] and Flicker2K [48] for training and test the model on several datasets including DIV2K validation set, Manga109 [14], Set5 [3] and Set14 [58]. Moreover, we evaluate our method for real-world image restoration task [56] to show its generalization capability, where the model is trained on ImageNet [9] and test on RealSR dataset [5].

Training details. Our implementation is mainly based on DDGAN [53] and DiT [44]. We adopt the same training configurations as DDGAN and DiT for all experiments. Training epochs are set as 250 (about 0.5M iterations with a batch size of 32) for face SISR and 300 (about 0.8M iterations with a batch size of 32) for general SISR (see Supplement for more hyperparameters and relevant details).

Evaluation metrics. For evaluation, we adopt two distortion-based metrics PSNR and SSIM [52], as well as perception-based metrics FID [19] and LPIPS [60]. Additionally, we adopt the identity similarity (denoted as "Deg.") [16] and consistency scores [46] (denoted as "Cons.") to measure the fidelity of outputs. Deg. means the face identity distance with angles between the restored image and ground-truth extracted by ArcFace [10]. Cons. measures the mean squared error (MSE) ($\times 10^{-5}$) between the down-sampled outputs and the LR image [46].

5.2. Comparison with state-of-the-art methods

Face SISR. Following IDM [15], we evaluate the proposed DTWSR on 100 face images from CelebA-HQ. Images are super-resolved from 16^2 to 128^2 pixels with $8\times$ upscaling. As illustrated in Table 1, our method shows promising quantitative results compared with SOTA. It achieves the lowest Cons. and Deg., indicating that DTWSR can preserve more face attributes to maintain the authenticity of outputs. The lowest FID score demonstrates its finer and more realistic detail generation. WFEN [33], trained with MSE loss merely, estimates the posterior mean effectively without considering data distribution, yielding over-smoothed images with poor perceptual quality. As visualized in Fig. 5, our method provides much more natural and realistic results with rich details such as exquisite skin, teeth and hair texture. In addition, our results are more similar to the ground-truth (*i.e.*, eyes, mouth and nose), without spatial distortion, leading to better objective fidelity.

General scene SISR. Following common practice [15, 30, 56] for fair comparison, We perform $4\times$ SR on general scene SISR datasets. In Table 2, we evaluate DTWSR on DIV2K validation set and compare the results with various prior arts, including regression-based and generative methods. Regression-based approaches (LIIF [8] and HAT [6]) yield higher PNSR and SSIM scores but worse LPIPS. our DTWSR shows better perception-distortion tradeoff [11] performance compared with other generative methods, with both higher reconstruction accuracy and better perceptual quality. The qualitative comparisons are in Fig. 6. Our method produces correct object structure with rich textural details (see the clearer animal fur). The regression-based HAT suffers from the typical over-smoothing issue. SRDiff [31], ResShift [56] and WGSR [30] are negatively affected by mis-alignment with LR condition, resulting in various artifacts (see the second and third rows).

Table 1. Quantitative comparison with several baselines on 16^2 to 128^2 face SISR. The best and second best results are highlighted in **bold** and underline.

Method	PSNR↑	SSIM↑	Cons.↓	Deg.↓	FID↓
FSRGAN [7]	23.01	0.62	33.8	-	-
SR3 [46]	23.04	0.65	2.68	58.99	70.82
DiWa [43]	23.34	0.67	-	-	-
IDM [15]	24.01	0.71	2.14	58.07	57.07
WFEN [33]	**25.53**	**0.77**	2.13	57.96	106.34
Ours	24.09	0.71	**0.50**	**53.85**	**56.77**

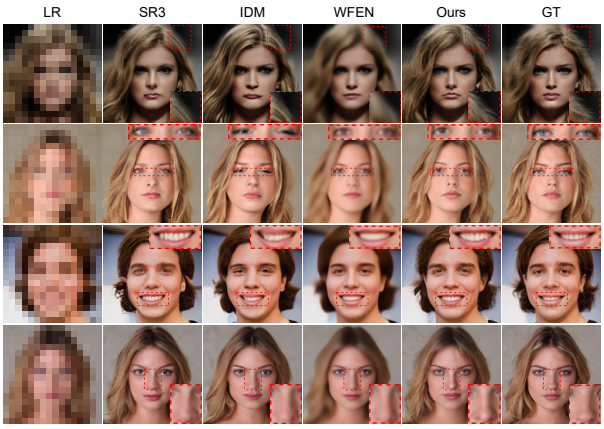


Figure 5. Qualitative comparison on $8\times$ SISR on CelebA-HQ. Our results not only maintain higher fidelity and more credible identities (eyes, mouth, *etc.*) close to the ground-truth, but also have finer textual details (skin, hair, *etc.*). Zoom in for best view.

We conducted more comparison on Manga109 [14], Set5 [3] and Set14 [58] as shown in Table 3. The proposed DTWSR outperforms generative methods on most metrics, which proves the effectiveness of our method further.

Large-magnification SISR. Here we explore DTWSR on large-magnification SISR. We test DTWSR on $12\times$ (16^2 to 196^2) and $16\times$ (16^2 to 256^2) face SISR and compare the results with SOTA method IDM [15][1] As shown in Fig. 7 and Table 4, IDM shows a significant drop in Cons. and FID un-

[1]We train IDM on $12\times$ and $16\times$ face SISR based the official code.

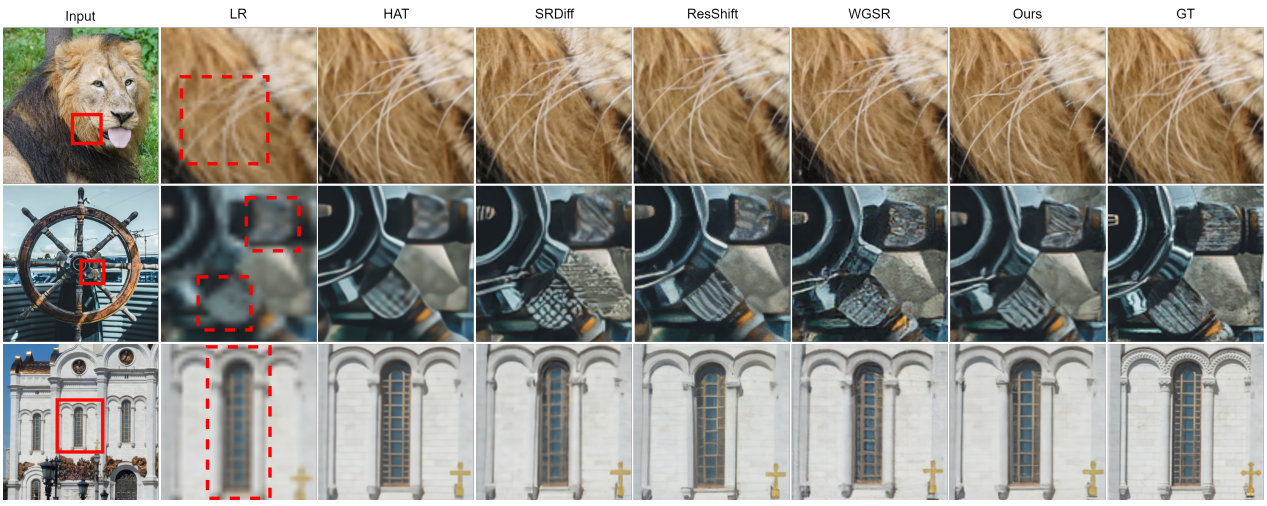


Figure 6. Qualitative comparison on 4× SISR on DIV2K [1]. The parts for detailed comparison are marked with red boxes in the images. Our results provides more credible details than other methods. Zoom in for best view.

Table 2. Quantitative comparison with several baselines on 4× general SISR. The best and second best results are highlighted in **bold** and underline among generative models.

Method	PSNR↑	SSIM↑	Cons.↓	LPIPS↓
Bicubic	26.70	0.77	17.86	0.409
LIIF [8]	29.29	0.82	0.820	0.132
HAT [6]	29.83	0.87	0.847	0.125
ESRGAN [49]	26.22	0.75	7.221	0.124
SRFlow [38]	27.09	0.76	-	0.120
SRDiff [31]	27.41	**0.79**	1.254	0.136
IDM [15]	27.59	0.78	-	-
DiWa [43]	28.09	0.78	-	0.104
ResDiff [47]	27.94	0.72	-	-
ResShift [56]	27.24	0.74	11.73	0.105
WGSR [30]	27.37	0.76	2.187	**0.096**
Ours	**28.18**	**0.79**	**1.151**	0.097

Table 3. Quantitative comparison on Manga109 [14], Set5 [3] and Set14 [58] dataset. The best and second best results are highlighted in **bold** and underline among generative models.

	Manga109 4×		Set14 4×		Set5 4×	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
SRDiff [31]	27.04	0.813	25.63	0.702	28.72	0.843
SR3 [46]	26.88	0.805	25.29	0.684	27.31	0.767
ResDiff [47]	27.76	0.832	26.19	0.718	29.32	**0.854**
ResShift [56]	26.91	0.824	25.11	0.682	28.54	0.817
WGSR [30]	26.59	0.823	25.28	0.644	27.65	0.781
Ours	**27.79**	**0.865**	**26.58**	**0.725**	**29.47**	0.846

der large-magnification SISR because of the less control on frequency layers. Our DTWSR maintain good performance on both fidelity and perceptual quality.

Real-world image restoration. We further explore the capability of DTWSR on real-world image restoration task. We train DTWSR on ImageNet training set [9] following the pipeline in ResShift [56] with the degradation model from RealESRGAN [50] adopted, and evaluate it on RealSR data[5]. A series of non-reference metrics, *e.g.*,

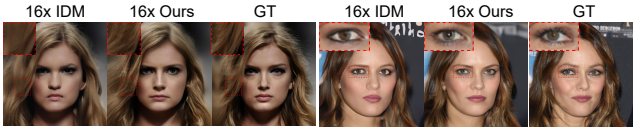


Figure 7. Qualitative comparison on 16× SISR (16^2 to 256^2) on CelebA-HQ [25]. Our results not only maintain higher fidelity but also have finer textual details. Zoom in for best view.

Table 4. Quantitative comparison with SOTA method [15] on large-magnification face SISR. Our results obtain better scores.

Method	PSNR↑	SSIM↑	Cons.↓	Deg.↓	FID↓
IDM-12×	23.21	0.65	8.16	55.44	60.58
Ours-12×	23.36	0.67	0.64	52.38	58.06
IDM-16×	23.15	0.65	9.57	56.53	63.99
Ours-16×	23.18	0.65	0.89	53.54	60.37

Table 5. Quantitative comparison with latest SOTA methods on real-world SISR. The best and second best results are highlighted in **bold** and underline.

Methods	MUSIQ↑	LIQE↑	NRQM↑	NIQE↓	PI↓
FlowIE [63]	56.83	2.44	4.84	5.68	5.53
ResShift [56]	56.14	2.80	6.20	7.34	5.55
SinSR [51]	61.45	3.19	**6.72**	5.76	4.49
Ours	**64.04**	**3.79**	6.70	**3.55**	**3.49**



Figure 8. Qualitative comparison on RealSR.

Table 6. Ablations of our approach for $8\times$ face SISR. Both components can introduce positive impact, while their fusion combines the strength of both components, resulting in the best scores.

Model	SISR on spectra	WSDT				Tokens↓	PSNR↑	SSIM↑	Cons.↓	Deg.↓	FID↓
		Pyramid tokenization	Dual decoder	LF residual	Attention mask						
Pixel-DiT	-	-	-	-	-	1040	23.94	0.714	1.77	53.47	61.372
Freq-DiT	√	√	-	-	-	704	24.02	0.715	0.582	55.17	60.352
DTWSR(a)	√	√	√	-	-	704	24.06	0.719	2.35	54.40	58.531
DTWSR(b)	√	√	√	√	-	704	24.09	0.718	0.549	56.19	60.008
DTWSR(c)	√	-	√	√	√	1040	23.63	0.694	1.147	57.30	66.051
DTWSR(ours)	√	√	√	√	√	704	24.09	0.719	0.503	53.85	56.771

Table 7. Ablation on the effect of interrelations among multi-scale frequency sub-bands.

	PSNR↑	SSIM↑	Cons.↓	Deg.↓	FID↓
w/o	23.99	0.711	1.154	55.00	61.670
w	24.09	0.719	0.503	53.85	56.771

MUSIQ [27], LIQE [61], NRQM [39], NIQE [42] and PI [4] are employed to justify the restoration quality, following common practice. As shown in Table 5, our method shows promising quality, surpassing existing methods on most metrics. As shown in Fig. 8, our method produces more natural results with clearer edges. FlowIE exhibits noticeable color drift, and SinSR introduces excessive noise.

5.3. Ablation studies

We conduct ablation studies on $8\times$ face SISR to evaluate the effectiveness of SISR on wavelet spectra and the proposed denoising network WSDT.

Effects of SISR on frequency domain. To show the effect of SR on wavelet frequency domain, we train a diffusion transformer model on pixel domain, named Pixel-DiT. Images are patchified spatially using a patch size of 4 [12] and LR tokens are concatenated to provide in-context conditioning. For fair comparison, we use the reconstruction loss as DTWSR with constraints on both pixel and frequency domains. As shown in Table 6, compared to Pixel-DiT, DTWSR enhances SR performance using only two-thirds the number of tokens.

Ablation on WSDT architecure. We perform ablation study on WSDT architecture, including the dual decoder design, LF Residual, attention mask and pyramid tokenization. We define Freq-DiT as our basic denoising network. It applies diffusion transformer on wavelet spectra but uses a unified decoder for all frequency sub-bands. In contrast, DTWSR(a) decodes LF and HF sub-bands independently by two decoders, without LF Residual considered. DTWSR(b) takes LF Residual into account but does not think over its influence on HF tokens. DTWSR(c) employs equal patch size (4) across frequency sub-bands.

As illustrated in Table 6, the distinct distribution in LF and HF affects each other in Freq-DiT, leading to worse results on Deg. and perceptual quality FID. When isolating decoders, DTWSR(a) shows better FID score straightaway. However, without aligning the relations between LF and HF sub-bands by LF Residual, it shows poor Cons.. DTWSR(b) Introduces LF Residual to promotes the alignment between LF and HFs, and shows improved Cons.. However, it resuffers the influence from LF to HF, leading to worse FID and Deg.. Therefore, we design attention mask M_{high} further in DTWSR(ours) to force the re-alignment between LF and HF sub-bands but avoid overwhelming of HF, leading to the best performance on both fidelity and perceptual quality.

Compared to DTWSR(c), DTWSR(ours) uses much less tokens, but achieves much better generation quality, which prove the advantage of our pyramid tokenization method.

Effects of correlation among frequencies. We employ a designed attention mask to artificially remove the interaction between different levels of HF sub-bands. As shown in Table 7, when the model ignores the interrelations among HF sub-bands, performance decreases across all metrics. The drop is particularly notable in FID, indicating significant degradation in image textural details.

6. Conclusion

In this work, we propose a diffusion transformer model based on image multi-level wavelet spectra, offering a novel solution for SISR. Our method integrates the strengths of diffusion models and transformers to capture the complex interrelations among multi-scale HF sub-bands. The pyramid tokenization promotes the learning of relationships between sub-bands for transformer. A dual-decoder transformer model is designed which separately processes the smooth contents in LF and the sparse HF details. The dedicated designed HDDec facilitates the exploration of correlations among frequency sub-bands, resulting in SR images with improved objective fidelity and perceptual quality. Extensive experiments demonstrate that our method achieves state-of-the-art performance on SISR across various SR magnification and diverse datasets. In the future, we intent to further explore the potential of multi-level wavelet spectra in promoting image generation.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 126–135, 2017. 6, 7

[2] Yuang Ai, Xiaoqiang Zhou, Huaibo Huang, Lei Zhang, and Ran He. Uncertainty-aware source-free adaptive image super-resolution with wavelet augmentation transformer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 3

[3] Marco Bevilacqua, Aline Roumy, Christine M. Guillemot, and Marie-Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Brit. Mach. Vis. Conf.*, 2012. 6, 7

[4] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *Eur. Conf. Comput. Vis. Worksh.*, pages 0–0, 2018. 8

[5] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Int. Conf. Comput. Vis.*, pages 3086–3095, 2019. 2, 6, 7

[6] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 22367–22377, 2023. 3, 6, 7

[7] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2492–2501, 2018. 2, 6

[8] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8628–8638, 2021. 6, 7

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 248–255. Ieee, 2009. 6, 7

[10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4690–4699, 2019. 6

[11] Xin Deng, Ren Yang, Mai Xu, and Pier Luigi Dragotti. Wavelet domain style transfer for an effective perception-distortion tradeoff in single image super-resolution. In *Int. Conf. Comput. Vis.*, pages 3076–3085, 2019. 1, 2, 3, 6

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Int. Conf. Learn. Represent.*, 2020. 4, 8

[13] Minghong Duan, Linhao Qu, Shaolei Liu, and Manning Wang. Local implicit wavelet transformer for arbitrary-scale super-resolution. In *Brit. Mach. Vis. Conf.*, 2024. 3

[14] Azuma Fujimoto, Toru Ogawa, Kazuyoshi Yamamoto, Yusuke Matsui, T. Yamasaki, and Kiyoharu Aizawa. Manga109 dataset and creation of metadata. *Proceedings of the 1st International Workshop on coMics ANalysis, Processing and Understanding*, 2016. 6, 7

[15] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10021–10030, 2023. 2, 6, 7

[16] Yuchao Gu, Xintao Wang, Liangbin Xie, Chao Dong, Gen Li, Ying Shan, and Ming-Ming Cheng. Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder. In *Eur. Conf. Comput. Vis.*, pages 126–143. Springer, 2022. 2, 6

[17] Tiantong Guo, Hojjat Seyed Mousavi, Tiep Huu Vu, and Vishal Monga. Deep wavelet prediction for image super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 104–113, 2017. 2

[18] Alfred Haar. Zur theorie der orthogonalen funktionensysteme.(zweite mitteilung). *Mathematische Annalen*, 71:38–53, 1912. 3

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Adv. Neural Inform. Process. Syst.*, 30, 2017. 6

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Adv. Neural Inform. Process. Syst.*, 33:6840–6851, 2020. 2

[21] Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In *Int. Conf. Comput. Vis.*, pages 1689–1697, 2017. 1, 2, 3

[22] Yi Huang, Jiancheng Huang, Jianzhuang Liu, Mingfu Yan, Yu Dong, Jiaxi Lyu, Chaoqi Chen, and Shifeng Chen. Wavedm: Wavelet-based diffusion models for image restoration. *IEEE Trans. Multimedia*, 26:2058–2073, 2024. 2, 3, 5

[23] Pranav Jeevan, Akella Srinidhi, Pasunuri Prathiba, and Amit Sethi. Wavemixsr: Resource-efficient neural network for image super-resolution. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024. 3

[24] Zahra Kadkhodaie, Florentin Guth, Stéphane Mallat, and Eero P Simoncelli. Learning multi-scale local conditional probability models of images. In *Int. Conf. Learn. Represent.*, 2023. 2

[25] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *Int. Conf. Learn. Represent.*, 2018. 6, 7

[26] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4401–4410, 2019. 6

[27] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Int. Conf. Comput. Vis.*, pages 5148–5157, 2021. 8

[28] Jinseok Kim and Tae-Kyun Kim. Arbitrary-scale image generation and upsampling using latent diffusion model and implicit neural decoder. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9202–9211, 2024. 2

[29] Jeongsoo Kim, Jongho Nang, and Junsuk Choe. Lmlt: Low-to-high multi-level vision transformer for image super-resolution. *arXiv preprint arXiv:2409.03516*, 2024. 3

[30] Cansu Korkmaz, A. Murat Tekalp, and Zafer Dogan. Training generative image super-resolution models by wavelet-domain losses enables better control of artifacts. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 1, 2, 3, 6, 7

[31] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022. 2, 6, 7

[32] Jinmin Li, Tao Dai, Mingyan Zhu, Bin Chen, Zhi Wang, and Shu-Tao Xia. Fsr: A general frequency-oriented framework to accelerate image super-resolution networks. In *AAAI*, pages 1343–1350, 2023. 2

[33] Wenjie Li, Heng Guo, Xuannan Liu, Kongming Liang, Jiani Hu, Zhanyu Ma, and Jun Guo. Efficient face super-resolution via wavelet-based feature enhancement network. In *ACM Int. Conf. Multimedia*, pages 4515–4523, 2024. 2, 3, 6

[34] Zhen Li, Zeng-Sheng Kuang, Zuo-Liang Zhu, Hong-Peng Wang, and Xiu-Li Shao. Wavelet-based texture reformation network for image super-resolution. *IEEE Trans. Image Process.*, 31:2647–2660, 2022. 3

[35] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Int. Conf. Comput. Vis. Worksh.*, pages 1833–1844, 2021. 3

[36] Jingyun Liang, Andreas Lugmayr, Kai Zhang, Martin Danelljan, Luc Van Gool, and Radu Timofte. Hierarchical conditional flow: A unified framework for image super-resolution and image rescaling. In *Int. Conf. Comput. Vis.*, pages 4076–4085, 2021. 2

[37] Jingyi Liu and Xiaomin Yang. Wtt: combining wavelet transform with transformer for remote sensing image super-resolution. *Machine Vision and Applications*, 2025. 3

[38] Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. Srflow: Learning the super-resolution space with normalizing flow. In *Eur. Conf. Comput. Vis.*, pages 715–732. Springer, 2020. 2, 7

[39] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, 2017. 8

[40] S.G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22:674 – 693, 1989. 2, 3

[41] Yunqi Miao, Jiankang Deng, and Jungong Han. Waveface: Authentic face restoration with efficient frequency recovery. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6583–6592, 2024. 2, 3, 5

[42] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Sign. Process. Letters*, 20(3):209–212, 2012. 8

[43] Brian B Moser, Stanislav Frolov, Federico Raue, Sebastian Palacio, and Andreas Dengel. Waving goodbye to low-res: A diffusion-wavelet approach for image super-resolution. In *2024 International Joint Conference on Neural Networks*, pages 1–8. IEEE, 2024. 2, 3, 6, 7

[44] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Int. Conf. Comput. Vis.*, pages 4195–4205, 2023. 5, 6

[45] Hao Phung, Quan Dao, and Anh Tran. Wavelet diffusion models are fast and scalable image generators. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10199–10208, 2023. 2

[46] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(4):4713–4726, 2022. 2, 6, 7

[47] Shuyao Shang, Zhengyang Shan, Guangxing Liu, LunQian Wang, XingHua Wang, Zekai Zhang, and Jinglin Zhang. Resdiff: Combining cnn and diffusion model for image super-resolution. In *AAAI*, pages 8975–8983, 2024. 2, 7

[48] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 114–125, 2017. 6

[49] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Eur. Conf. Comput. Vis. Worksh.*, pages 0–0, 2018. 2, 7

[50] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Int. Conf. Comput. Vis.*, pages 1905–1914, 2021. 2, 7

[51] Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. Sinsr: diffusion-based image super-resolution in a single step. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 25796–25805, 2024. 2, 7

[52] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4): 600–612, 2004. 6

[53] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. In *Int. Conf. Learn. Represent.*, 2022. 2, 5, 6

[54] Jingwei Xin, Jie Li, Xinrui Jiang, Nannan Wang, Heng Huang, and Xinbo Gao. Wavelet-based dual recursive network for image super-resolution. 33(2):707–720, 2020. 2

[55] Yutao Yuan and Chun Yuan. Efficient conditional diffusion model with probability flow sampling for image super-resolution. In *AAAI*, pages 6862–6870, 2024. 2

[56] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Efficient diffusion model for image restoration by residual shifting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024. 2, 6, 7

[57] Syed Waqas Zamir, Aditya Arora, Salman Khan, and Munawar Hayat. Restormer: Efficient transformer for high-resolution image restoration. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5728–5739, 2022. 3

[58] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces*, 2010. 6, 7

[59] Dafeng Zhang, Feiyu Huang, Shizhuo Liu, Xiaobing Wang, and Zhezhu Jin. Swinfir: Revisiting the swinir with fast fourier convolution and improved training for image super-resolution. *arXiv preprint arXiv:2208.11247*, 2022. 3

[60] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 586–595, 2018. 6

[61] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 14071–14081, 2023. 8

[62] Yuehan Zhang, Bo Ji, Jia Hao, and Angela Yao. Perception-distortion balanced admm optimization for single-image super-resolution. In *Eur. Conf. Comput. Vis.*, pages 5926–5936, 2022. 3

[63] Yixuan Zhu, Wenliang Zhao, Ao Li, Yansong Tang, Jie Zhou, and Jiwen Lu. Flowie: Efficient image enhancement via rectified flow. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13–22, 2024. 2, 7

[64] Wenbin Zou, Liang Chen, Yi Wu, Yunchen Zhang, Yuxiang Xu, and Jun Shao. Joint wavelet sub-bands guided network for single image super-resolution. *IEEE Trans. Multimedia*, 25:4623–4637, 2022. 3