WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 4, Number 3, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v4n3.25
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Fkdiff: An Efficient Diffusion Model for Image Super-Resolution With Fourier Frequency Domain Transformation and Knowledge Distillation

Yu Chen

School of Artificial Intelligence and Data Science, Hebei University of Technology, Tianjin, Tianjin, 300000, China

ABSTRACT

Image super-resolution (SR) techniques play a crucial role in various applications such as image restoration, medical imaging, surveillance, and remote sensing. Traditional methods often employ interpolation algorithms to upscale images, resulting in artifacts and reduced perceptual quality. Recent advancements in diffusion models (DM) have shown promising results in image generation tasks but are hindered by computational complexity, particularly in resource-constrained environments. By leveraging low-resolution images as prior information and operating in the frequency domain, FKdiff achieves enhanced computational efficiency and preserves high-frequency details effectively. The proposed method integrates a progressive hexagonal knowledge distillation (PHexKD) approach, ensuring lightweight model deployment without compromising performance. Experimental results demonstrate that FKdiff outperforms existing methods in terms of efficiency and effectiveness while a small amount of image generation quality is lost.

KEYWORDS

Image super-resolution; Diffusion Models; Knowledge Distillation; Computational efficiency

1. INTRODUCTION

Image super-resolution technology finds wide applications in the restoration of old photographs, medical image processing, enhancement of surveillance footage, and the enlargement of remote sensing images. With the proliferation of high-resolution devices, existing resources have not kept pace with this technological advancement, creating a pressing need for methods to bring current resources up to par with modern equipment. Traditional interpolation algorithms (such as nearest, bilinear, bicubic, lanczos, sinc, spline, etc.) often result in images with noticeable defects like blurriness, jagged edges, and artifacts, leading to unsatisfactory visual experiences. Although current super-resolution models (such as GANs, DMs and regression-based methods) achieve good results, their high computational cost makes them challenging to deploy on edge devices.

Diffusion Probabilistic Models (DPMs) [1] have recently emerged as a powerful class of generative models, showing impressive results in image generation. However, achieving satisfactory outcomes comes at the cost of large model sizes, which makes their deployment on resource-constrained platforms extremely difficult. Current research on diffusion models primarily focuses on accelerating model sampling speeds. While significant strides have been made in this area, the aspect of model lightweighting has been largely overlooked, especially in the context of image super-resolution tasks.

Given that low-resolution images can serve as a reference, there is no need to regenerate the entire image for super-resolution tasks. Existing studies often neglect this crucial factor, resulting in

unnecessary redundancy in models. Additionally, processing images in the spatial domain requires substantial MAC operations, whereas handling images in the frequency domain via Fourier transform is more efficient and better at preserving high-frequency details, leading to more vivid images. Knowledge distillation [2], a widely used lightweighting method, also shows unique advantages in this field. Therefore, this paper proposes a lightweight diffusion model FKdiff for image super-resolution in the frequency domain, which incorporates progressive hexagonal knowledge distillation (PHexKD) and utilizes low-resolution images to provide prior information.

2. RELATED WORK

2.1. Efficient Diffusion Models

Efficient and cost-effective diffusion models have emerged as a significant research focus due to their ability to generate high-quality super-resolution (SR) images. However, these models are predominantly based on pixel-domain diffusion processes, leading to high computational costs and inefficiencies. LDM [3] address this issue by transforming input data into a low-dimensional latent space, significantly reducing resource usage without compromising model performance. Nevertheless, this method loses some details during data compression. To address this, researchers have proposed compressing data into the Wavelet Target Domain, such as DiWa [4], WSGM [5], and ResDiff [6]. This approach reduces the spatial size of an image by four times and retains high-frequency details. Stable Diffusion XL [7] is a successful two-stage cascaded diffusion model that balances speed and accuracy through the combination of Base and Refiner models.

Unlike the aforementioned generation tasks, image super-resolution tasks can utilize low-quality images as prior information to guide the generation of high-resolution images, thereby improving inference efficiency. LDM and DiffIR [8] effectively accelerate model inference speed by incorporating prior information. Notably, DiffIR achieves state-of-the-art (SOTA) performance among current latent diffusion models. This model abandons the traditional UNet denoising network, instead using linear layers for denoising guided by extracted prior features, and ultimately feeds the results into a pre-trained UNet for image recovery. This approach significantly reduces model parameters and computational load while maintaining accuracy. However, the image recovery process still occurs in the spatial domain, necessitating substantial computational resources and failing to fully extract global information. SinSR [9] employs a progressive knowledge distillation method to distill the inference process step by step, achieving one-step inference. However, its generated image quality is not ideal, with issues of incomplete denoising and artifacts.

2.2. Attention in the Frequency Domain

In digital image processing, the Fourier frequency domain represents an image using a set of sine waves, with each wave representing a part of the image at different intensity levels. The frequency domain is an effective method for understanding images with repetitive or periodic patterns. Compared to traditional spatial domain techniques, it more effectively captures geometric structures that are difficult to extract. By capturing intensity variations in an image, the frequency domain can identify different regions related to objects. Each frequency in the frequency domain is determined by all pixels in the spatial domain. High frequencies correspond to significant intensity changes over short distances between pixels (such as edges). Thus, focusing on the frequency domain can be considered a form of global attention. Komodakis et al. [10] pointed out that spatial domain attention primarily affects local areas in the input feature map, which may not be sufficient to capture the global structure. In contrast, frequency domain attention is particularly useful for identifying global information or geometric structures in feature maps. Kong et al. [11] highlighted that in the spatial domain, Transformers require extensive convolutions, leading to high computational load and loss of

high-frequency details. In the frequency domain, these convolutions can be replaced with element-wise multiplications, optimizing computation and preserving information.

2.3. Knowledge Distillation in SISR

In 2015, Hinton et al. [12] summarized and popularized the concept of Knowledge Distillation(KD), garnering widespread attention. In KD, a student model benefits from various forms of guidance from a teacher model to achieve enhanced performance. However, most existing KD research focuses on high-level vision tasks, such as image classification [13] and semantic segmentation [14], with less attention on low-level vision tasks like image denoising and super-resolution. A representative work in low-level vision KD is by Wang et al. [15]. They proposed a cooperative distillation method, successfully compressing several classical network architectures. FAKD [16] leverages feature map correlation to improve distillation performance, while PISR [17] uses ground truth high-resolution images as privileged information along with feature distillation to enhance the performance of super-resolution networks. However, these distillation methods are performed in the spatial domain, necessitating substantial computational operations and struggling to extract global and high-frequency information compared to frequency domain distillation. Pham et al. [18] proposed a frequency domain KD method for super-resolution, converting feature maps to the frequency domain for key information screening before converting them back to the spatial domain for comparison. In this paper, we proposed a new cooperative KD method in the frequency domain for super-resolution tasks, optimizing the computational process and reducing both model size and inference time.

3. METHODOLOGY

In recent years, diffusion models have achieved significant progress in image generation tasks, demonstrating robust probability density estimation capabilities. However, traditional diffusion models require numerous denoising networks to infer the final result, leading to prolonged inference times and large model sizes. Inspired by latent diffusion models and DiffIR, we address this issue by compressing low-quality (LQ) images to provide prior information to the denoising network. This approach accelerates the denoising process while achieving superior results.

Given that Vision Transformers (ViT) are popular for restoring compressed features estimated by latent diffusion models, they inevitably involve extensive multi-add operations. Therefore, we propose FASRer, which is composed of two modules: CFGA and FFFN.

The CFGA module employs Fast Fourier Transform (FFT) to shift the traditional Transformer attention mechanism to the frequency domain. This not only facilitates the extraction of global and high-frequency information but also significantly reduces computational load. Our proposed FFFN module introduces learnable frequency filters to screen useful frequency information.

Additionally, we propose a Progressive Hexagonal Knowledge Distillation (PHexKD) method, which adopts a hexagonal architecture with two lightweight student models. Through progressive distillation, this method narrows the gap between the large teacher model and the student models, achieving improved results. The following sections provide detailed explanations of each component.

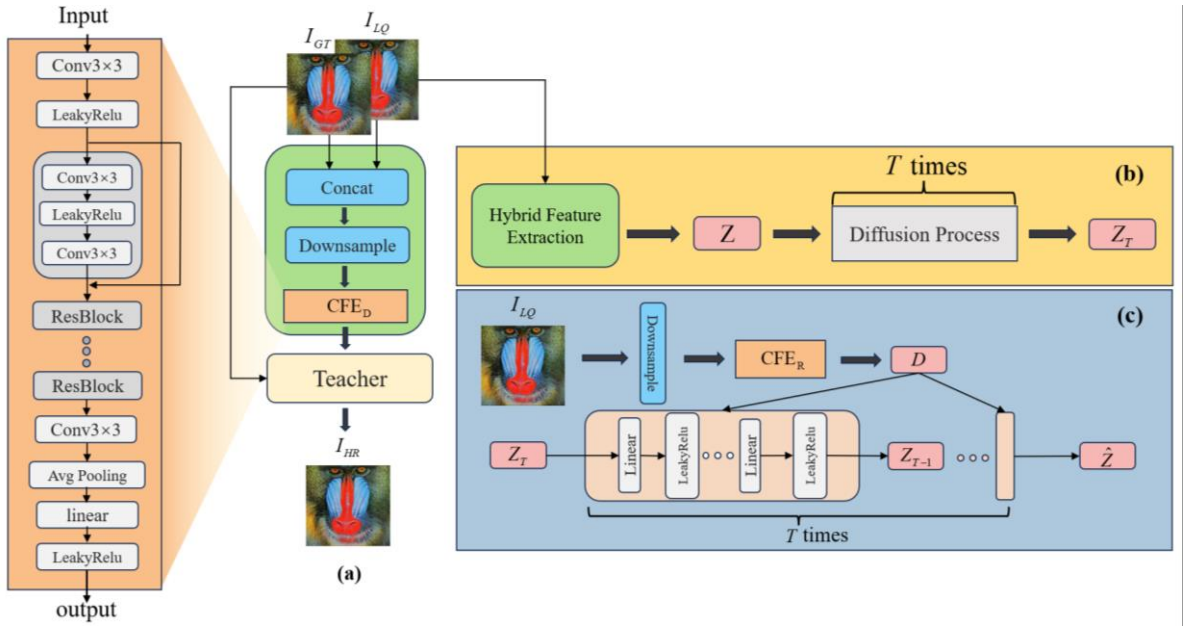

Figure 1. This diagram illustrates the pre-training stage of the teacher model. Part (a) shows the collaborative training process of the teacher model and CFE_D. Note: After training CFE_D, its parameters are locked and used for hybrid feature extraction in part (b). Part (b) depicts the diffusion process where the low-quality (LQ) image is compressed by CFE_D into a compact prior feature Z, which is then used in the diffusion process to obtain the prior feature Z_T. Part (c) illustrates the denoising process, where Z_T is used to estimate $\hat{Z}$, which is then used for the subsequent image reconstruction network.

3.1. Efficient Diffusion Model for Image Super Resolution

Due to the flexible design of network structures in diffusion models [1], and inspired by latent diffusion models [3] and DiffIR [8], we utilize compressed features as the target for denoising, guiding and accelerating our denoising process. For the diffusion process, given an input image x_0, by progressively adding Gaussian noise T times, we obtain $x_T \sim \mathcal{N}(0,1)$. The intermediate process can be described as:

$$q(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I\right)$$

Where x_t is the noisy picture at step t, β_t is a predefined scaling factor and $\mathcal{N}$ means Gaussian distribution. The transition from x_0 tox_t can also be represented as:

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

Where $\alpha_t = 1-\beta_t$, $\bar{\alpha}_t = \prod_{i=0}^{t}\alpha_i$. For the denoising process, DM samples Gaussian noise $x_T \sim \mathcal{N}(0,1)$ and iteratively denoises it, akin to simulated annealing, until it converges x_T to a high-resolution imagex_0:

$$p(x_{t-1}|x_t, x_0) = N(x_{t-1}; \mu_t(x_t, x_0), \sigma_t^2 I)$$

Where the mean is $\mu_t(x_t, x_0) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \varepsilon\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\right)$, $\varepsilon$is the unknown noise, and variance is $\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$. To obtain x_0, our goal becomes estimating the noise ε at each time step t. DM uses a denoising network $\varepsilon_\theta(x_t, t)$ to estimate the noise. During the diffusion process, given the distribution of $x_t$each step, the objective of training a denoising network is:

$$\nabla_\theta \left\| \varepsilon - \varepsilon_\theta \left(\sqrt{\bar{\alpha}_t} x_0 + \varepsilon \sqrt{1 - \bar{\alpha}_t}, t \right) \right\|_2^2$$

Where ε_θ denotes the denoising network at step t, $\varepsilon \sim \mathcal{N}(0,1)$.

Before formally training our network, we first pretrain the Condensed Feature Extractor (CFE) with the Teacher Model to obtain matched CFE_D. This step requires low-quality (LQ) images and ground truth (GT) images as inputs, as shown in part (a) of Figure 1. Note that in this part, the teacher model should be appropriately modified to correctly configure the compressed prior features $Z \in \mathbb{R}^{C'}$ obtained by CFE_D for reconstruction network training, this process can be represented as:

$$Z = CFE_D(Downsample(Concat(I_{GT},\ I_{LQ})))$$

$$I_{HR} = Teacher(I_{LQ}, Z)$$

$$\mathcal{L}_{L1}(I_{GT},\ I_{HR}) = \frac{1}{N} \sum_{i=1}^{N} ||I_{GT} - I_{HR}||_1$$

Where Z denotes the compressed prior features, I_{HR} represents the high-resolution image generated by the teacher model with compressed features, $\mathcal{L}_{L1}$ denotes the L1 loss, and N represents the number of pixels (H×W×C). The parameters of the trained CFE_D will be locked and used for the mixed feature extraction of the diffusion model. After the diffusion process, we obtain $Z_T \in \mathbb{R}^{C'}$, as shown in part (b) of Figure 1, which can be represented as:

$$q(Z_T \mid Z) = \mathrm{N}\left(Z_T; \sqrt{\bar{\alpha}_T} Z, (1 - \bar{\alpha}_T)\mathrm{I}\right)$$

The obtained Z_T will serve as the denoising target for the reverse process. During the denoising process, we use CFE_R to generate the guiding vector $D \in \mathbb{R}^{C'}$:

$$D = CFE_R(Downsample(I_{LQ}))$$

Next, we can utilize the guiding vector D to guide the denoising and estimate the prior features $\hat{Z}$. Finally, the estimated prior features $\hat{Z}$ are connected to the reconstruction network for image reconstruction. The process to obtain $\hat{Z}$ can be represented as:

$$\hat{Z}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\hat{Z}_t - \varepsilon \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \right)$$

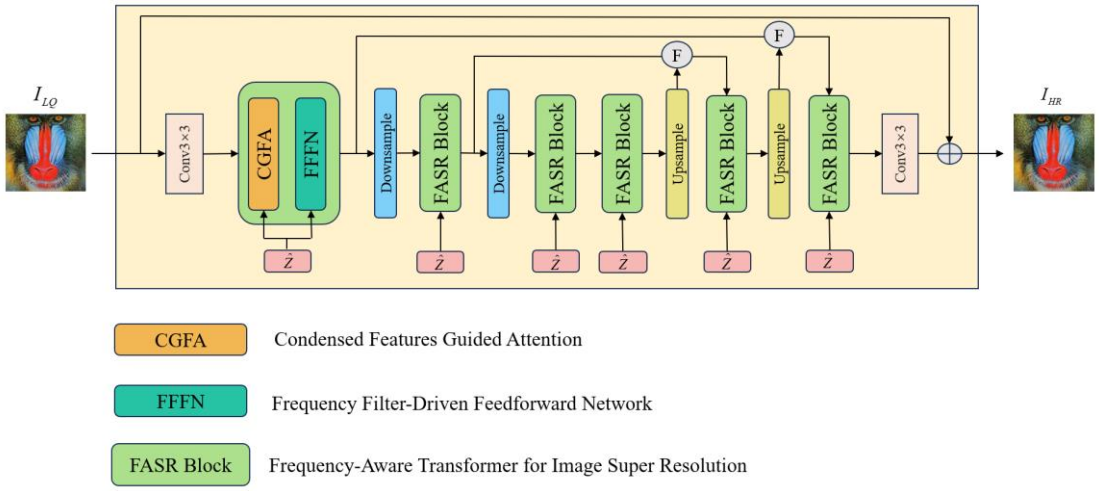

Figure 2. The Frequency-Aware Transformer for Image Super Resolution (FASRer) consists of two main components: the Condensed Features Guided Attention (CFGA) and the Frequency Filter-Driven Feedforward Network (FFFN). This architecture includes several skip connections, with "F" in the figure representing feature fusion.

3.2. Frequency-Aware Transformer for Image Super Resolution (FASRer)

We employ the Frequency-Aware Transformer (FASRer) as our image reconstruction network, serving as our student network. This network takes the prior features estimated by the diffusion model and the low-quality (LQ) image as inputs. It extracts features through meticulously designed modules: the Condensed Features Guided Attention (CFGA) and the Frequency Filter-Driven Feedforward Network (FFFN). The overall architecture of the network is illustrated in Figure 2. The following sections will provide a detailed explanation of each module.

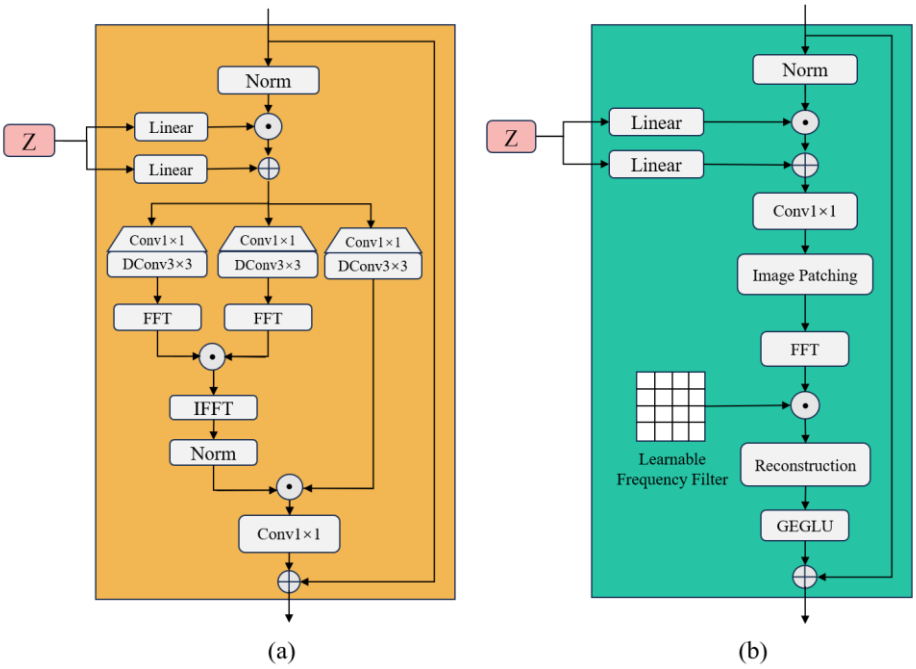

Figure 3. (a) illustrates CFGA mechanism, which utilizes the condensed prior features, and incorporates the Fast Fourier Transform (FFT) for processing. (b) depicts the structure of FFFN.

3.2.1. Condensed Features Guided Attention (CFGA)

We propose the Condensed Features Guided Attention (CFGA) module, as illustrated in Figure 3(a). For an input feature map X with resolution H×W×C (where H is height, W represents width, and C

is the number of channels), and a given compressed feature$Z$with C'channels, a linear transformation L is applied to Z to match the C channels of the input feature map. The transformed Z is then element-wise multiplied and added to the input feature map for feature fusion. Then, the fused feature map is split into three parts via linear transformations W_q, W_k and W_v, resulting in F_k, F_q and F_v:

$$Z' = L(Z)$$

$$X' = Concat(X, Z')$$

$$F_i = W_i X', i \in \{k, q, v\}$$

In Vision Transformers (ViT), F_k,$F_q$andF_v are often segmented into patches [11; 19] to reduce the calculation cost and maintain its spatial structure to a certain extent. These patches are represented as $\{k_i\}_{i=1}^{N}$, $\{q_i\}_{i=1}^{N}$, $\{v_i\}_{i=1}^{N}$. After reshaping, the concatenated patches are denoted as K, Q, V, respectively, and the scaled dot-product attention can be formulated as:

$$V_{att} = \text{softmax}\left(\frac{QK^T}{\sqrt{CH_pW_p}}\right)V$$

Where H_p and W_p are the height and width of each patch, respectively. Each element of the attention matrix can be computed by:

$$(QK^T)_{ij} = \langle q_i, k_j \rangle$$

Where q_i, k_j are the i-th and j-th vectors of $\{q_i\}_{i=1}^{N}$, $\{v_i\}_{i=1}^{N}$.

Inspired by [11], we apply the Fast Fourier Transform (FFT) to F_k and F_q to efficiently compute the attention:

$$A = \mathcal{F}^{-1}\left(\mathcal{F}(F_q) \odot \overline{\mathcal{F}(F_k)}\right)$$

Where $\mathcal{F}(\quad)$and$\mathcal{F}^{-1}(\quad)$ denote the FFT and inverse FFT, respectively, and $\overline{\mathcal{F}(\quad)}$ represents the conjugate transpose of $\mathcal{F}(\quad)$.Thus, the attention values are computed as follows:

$$V_{att} = Norm(A)F_v$$

Finally, the output of the CFGA can be obtained by:

$$X_{CFGA} = X + \text{Conv}_{1\times1}(V_{att})$$

3.2.2. Frequency Filter-Driven Feedforward Network (FFFN)

In the Frequency Filter-Driven Feedforward Network (FFFN), we utilize compressed features and frequency filters within the FFN network. The compressed features guide the FFN network, providing information from the original LQ image and reducing information loss as the network deepens. Since convolution operations tend to restore low and mid-frequency information, frequency filters allow us to select important information, such as high-frequency details. The overall process can be described as follows:

$$Z' = L(Z)$$
$$X_F = Concat(Norm(X_{CFGA}), Z')$$

$$X_F^f = \mathcal{F}\left(\mathcal{P}\left(Conv_{1\times1}(X_F)\right)\right)$$

$$X_F^{'} = \mathcal{F}^{-1}\left(WX_F^f\right)$$

$$X_{FFFN} = \mathcal{G}(P^{-1}(X_F^{'})) + X_{CFGA}$$

Where Z represents the compressed prior features, $L()$denotes the linear transformation, $\mathcal{F}(\ \ )$ and $\mathcal{F}^{-1}(\ \ )$represent the FFT and reverse FFT, respectively. $Norm()$ denotes layer normalization. $\mathcal{P}()$means patching process which divides feature maps into many patches. $\mathcal{P}^{-1}()$refer to reshaping the feature map into patches and reconstructing it to the original size, while $\mathcal{G}()$is the activation function.

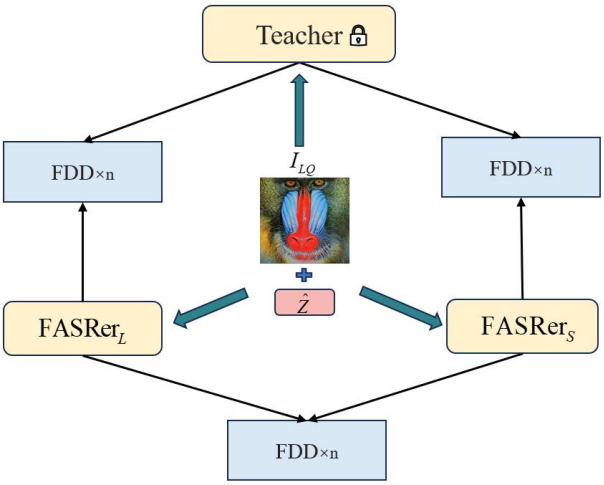

Figure 4. The figure illustrates the Progressive Hexagonal Knowledge Distillation (PHexKD) network. This network takes LQ images and compressed features estimated by the diffusion model as inputs. The architecture comprises a teacher model, a smaller student model FASRer$_S$, and a larger student model FASRer$_L$, along with numerous Frequency Discriminative Distiller (FDD) blocks.

3.3. Progressive Hexagonal Knowledge Distillation (PHexKD)

3.3.1. Hexagonal Architecture

The knowledge distillation framework adopts a hexagonal shape, as illustrated in Figure 4. The network utilizes LQ images and the compressed features$\hat{Z}$estimated by the diffusion model as inputs. It includes both a smaller student model FASRer$_S$ and a larger student model FASRer$_L$, each with different parameter counts but similar overall structures. These models, along with the teacher model, transfer knowledge through Frequency Discriminative Distiller (FDD) blocks. Each FDD block corresponds to a specific layer between the two networks. Through the FDD blocks, we can reduce the differences between feature maps of corresponding layers in the two networks while preserving important frequency information during knowledge transfer. The distillation loss between Network 1 and Network 2 can be expressed as:

$$\mathcal{L}_{KD}^{1,2} = \mathcal{L}_{FDD}^{1,2} + \mathcal{L}_{L1}(I_{HR}^1, I_{HR}^2)$$

Where $\mathcal{L}_{KD}^{1,2}$ represents the distillation loss, $\mathcal{L}_{FDD}^{1,2}$ denotes the distillation loss for all FDDs between Network 1 and Network 2, and $\mathcal{L}_{L1}$ is the L1 loss. I_{HR}^1 and I_{HR}^2 are the high-resolution images

reconstructed by Network 1 and Network 2, respectively. Therefore, the total loss for the distillation network can be expressed as:

$$\mathcal{L}_{pix} = \sum_{i} \mathcal{L}_{L1}\left(I_{GT}, I_{HR}^{i}\right), i \in \{s, l, t\}$$

$$\mathcal{L}_{KD} = \mathcal{L}_{KD}^{s,t} + \mathcal{L}_{KD}^{l,t} + \mathcal{L}_{KD}^{s,l} + \mathcal{L}_{pix}$$

Where $\mathcal{L}_{pix}$ is the defined pixel loss, I_{GT} is the ground truth high-resolution image, and I_{HR}^{i} is the high-resolution image generated by Network i. The set $i \in \{s, l, t\}$ represents all networks involved, where s denotes FASRer$_S$, l denotes FASRer$_L$, and$t$ denotes the teacher network.

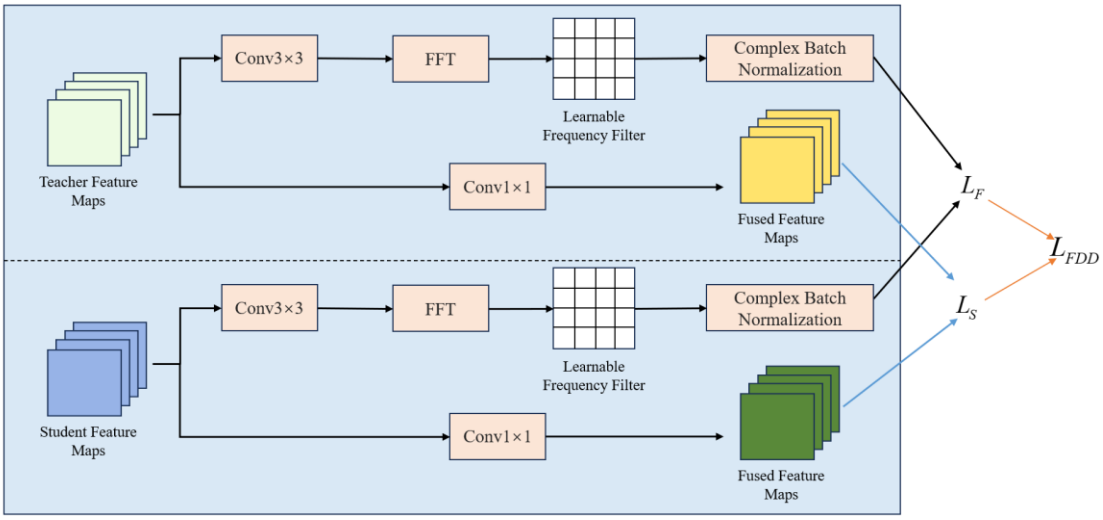

Figure 5. The Frequency Discriminative Distiller (FDD) is composed of two branches: the upper branch represents the teacher network, and the lower branch represents the student network. Each branch takes a feature map from a specific layer of the network as input and outputs both frequency domain feature maps and fused feature maps.

3.3.2. Frequency Discriminative Distiller (FDD)

The FDD module aims to minimize the differences between the feature maps of Network 1 and Network 2 by deploying FDD blocks at specific layers of the teacher and student networks. The structure of the FDD is illustrated in Figure 5. It consists of two branches, with the upper part comparing the differences in the frequency domain between the two networks' feature maps. In this part, the feature maps are first further dimensionally reduced and extracted through convolution, then transformed into the frequency domain via Fast Fourier Transform (FFT). After that, a Learnable Frequency Filter is applied to filter out useful frequency band information. Finally, the filtered spectral feature maps undergo complex domain batch normalization:

$$i \in \{s, t\}$$

$$X_{f_1}^{i} = Conv_{3\times3}\left(X^{i}\right)$$

$$X_{f_2}^{i} = \mathrm{W}\mathcal{F}\left(X_{f_1}^{i}\right)$$

$$X_{fout}^{i} = \mathcal{C}(X_{f_2}^{i})$$

Where $i \in \{s,t\}$ represents the set of network types, s represents the student network, and t represents the teacher network. W denotes the weight matrix. $\mathcal{C}(\ )$ represents complex domain batch normalization. $X_{f_1}^i$, $X_{f_2}^i$, $X_{f_{out}}^i$ represent the first layer, second layer, and the output feature maps of the frequency domain information extraction branch of network i, respectively. The lower part of the FDD uses $Conv_{1\times1}$ to fuse the feature maps, comparing the spatial domain differences between the feature maps:

$$X_S^i = Conv_{1\times1}(X^i)$$

Where X_S^i represents the spatial domain feature maps of network i. To better evaluate the difference between two frequency domain feature maps, we introduce a frequency domain loss [20]. For a given input $X \in \mathbb{R}^{H\times W\times C}$, after Discrete Fourier Transform (DFT), X transforms into $X \in \mathbb{R}^{U\times V\times C}$:

$$\mathcal{F}(X) = X_{u,v} = \frac{1}{\sqrt{HW}} \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} x_{h,w} e^{-i2\pi(u\frac{h}{H}+v\frac{w}{W})}$$

Where $\mathcal{F}(\ )$ represents Discrete Fourier Transform. After DFT, the amplitude $\left|X_{u,v}\right|$of $X_{u,v}$ can be expressed as:

$$\left|X_{u,v}\right| = \sqrt{\mathcal{R}\left(X_{u,v}\right)^2 + \mathcal{I}\left(X_{u,v}\right)^2}$$

Where $\mathcal{R}\left(X_{u,v}\right)$ and$\mathcal{I}\left(X_{u,v}\right)$ represent the real and imaginary parts of $X_{u,v}$, respectively. The phase $\angle X_{u,v}$ can be expressed as:

$$\angle X_{u,v} = \text{atan2}\left(\mathcal{I}\left(X_{u,v}\right), \mathcal{R}\left(X_{u,v}\right)\right)$$

After obtaining $X_{f_{out}}^i$, we can compute the amplitude loss and phase loss:

$$\mathcal{L}_{\mathcal{F},amp} = \frac{2}{UV} \sum_{u=0}^{U/2-1} \sum_{v=0}^{V-1} \left|\left|X_{f_{out}}^t\right| - \left|X_{f_{out}}^s\right|\right|$$

$$\mathcal{L}_{\mathcal{F},ph} = \frac{2}{UV} \sum_{u=0}^{U/2-1} \sum_{v=0}^{V-1} \left|\angle X_{f_{out}}^t - \angle X_{f_{out}}^s\right|$$

Where $\mathcal{L}_{\mathcal{F},amp}$, $\mathcal{L}_{\mathcal{F},ph}$ are used to compute the amplitude and phase discrepancies between the frequency domain feature maps of the teacher and student models. U and V denote the height and width of the images. Notably, due to the symmetry in the frequency domain, half of the components are redundant. Therefore, we set $u$as U/2-1 to reduce the computation. The frequency domain loss can be expressed as:

$$\mathcal{L}_{\mathcal{F}} = \eta\mathcal{L}_{\mathcal{F},amp} + (1-\eta)\mathcal{L}_{\mathcal{F},ph}$$

Where η is weight parameter. By using X_S^i, we can compute the spatial domain loss $\mathcal{L}_S$ to measure the differences between the spatial domain feature maps:

$$\mathcal{L}_{\mathcal{S}} = \mathcal{L}_{L1}(X_{\mathcal{S}}^{t}, X_{\mathcal{S}}^{s})$$

Therefore, the total loss function of FDD can be expressed as:

$$\mathcal{L}_{FDD} = \lambda \mathcal{L}_{\mathcal{F}} + (1-\lambda)\mathcal{L}_{\mathcal{S}}$$

Where λ weight parameter.

4. EXPERIMENTS

4.1. Experiment Settings

The experiments were trained on a single A100 GPU. DF2K [21; 22] was used as the training set. In the pre-training phase, the learning rate was set to 0.0002, with the Adam optimizer and a batch size of 9. During the knowledge distillation phase, the learning rate was set to 0.0001, with the Adam optimizer and a batch size of 5.

In the distillation part, our teacher network used DiffIRS1. The student network adopted a 3-layer Encoder-decoder network. The Transformer Block for Network1 was configured as [8, 2, 2, 2], and for Network2 as [6, 1, 1, 2]. The CFE channel number was set to 64. In the diffusion model part, timesteps were set to 4.

Table 1. Quantitative evaluation of single image super-resolution on different benchmark datasets. The best and second-best performances are highlighted in red and blue, respectively. The bottom four methods, marked in gray, utilize the diffusion model.

Method	Set14 [23]			DIV2K100 [21]			Urban100 [24]			BSD100 [25]			Manga109 [26]		
	PSNR ↑	SSIM↑	LPIPS↓	PSNR ↑	SSIM↑	LPIPS↓	PSNR ↑	SSIM↑	LPIPS↓	PSNR ↑	SSIM↑	LPIPS↓	PSNR ↑	SSIM↑	LPIPS↓
SRGAN [27]	27.14	0.7349	0.1489	28.74	0.7933	0.1416	24.86	0.7475	0.1541	26.22	0.6856	0.1996	28.34	0.8682	0.0732
ESRGAN [28]	26.29	0.6978	0.1337	28.17	0.7759	0.1154	24.35	0.7355	0.1230	24.35	0.7325	0.1615	28.41	0.8595	0.0649
Beby-GAN [29]	26.89	0.7267	0.1231	28.62	0.7899	0.1021	25.23	0.7623	0.1094	25.78	0.6765	0.1512	29.19	0.8754	0.0529
CAL-GAN [30]	27.33	0.7355	0.1332	28.95	0.7897	0.1075	25.34	0.7624	0.1170	26.27	0.7620	0.1681	29.21	0.8675	0.0699
SRDiff [31]	25.36	0.7360	0.1422	27.20	0.7883	0.1292	23.69	0.7550	0.1410	24.46	0.6790	0.1998	27.30	0.8710	0.0640
DiffIR [8]	27.06	0.7343	0.1195	29.11	0.7958	0.0881	25.96	0.7786	0.1016	26.23	0.6832	0.1490	30.22	0.8913	0.0477
SinSR [9]	23.98	0.6743	0.1772	26.41	0.7615	0.1905	22.45	0.7032	0.1755	23.70	0.6359	0.2137	26.06	0.8459	0.1087
Ours	26.77	0.7280	0.1238	29.05	0.7955	0.1036	25.33	0.7614	0.1169	26.11	0.6769	0.1610	29.45	0.8829	0.0555

4.2. Evaluation

We evaluated our model on various single-image super-resolution (SISR) benchmark datasets, comparing it against SOTA methods known for high perceptual quality. The evaluation metrics used were PSNR, SSIM, and LPIPS [32]. The test results are presented in Table 1. To better illustrate the

quality of our method's output, we visualized the generated images and compared them with those produced by other SOTA perceptual quality-based methods, as shown in Figure 6.

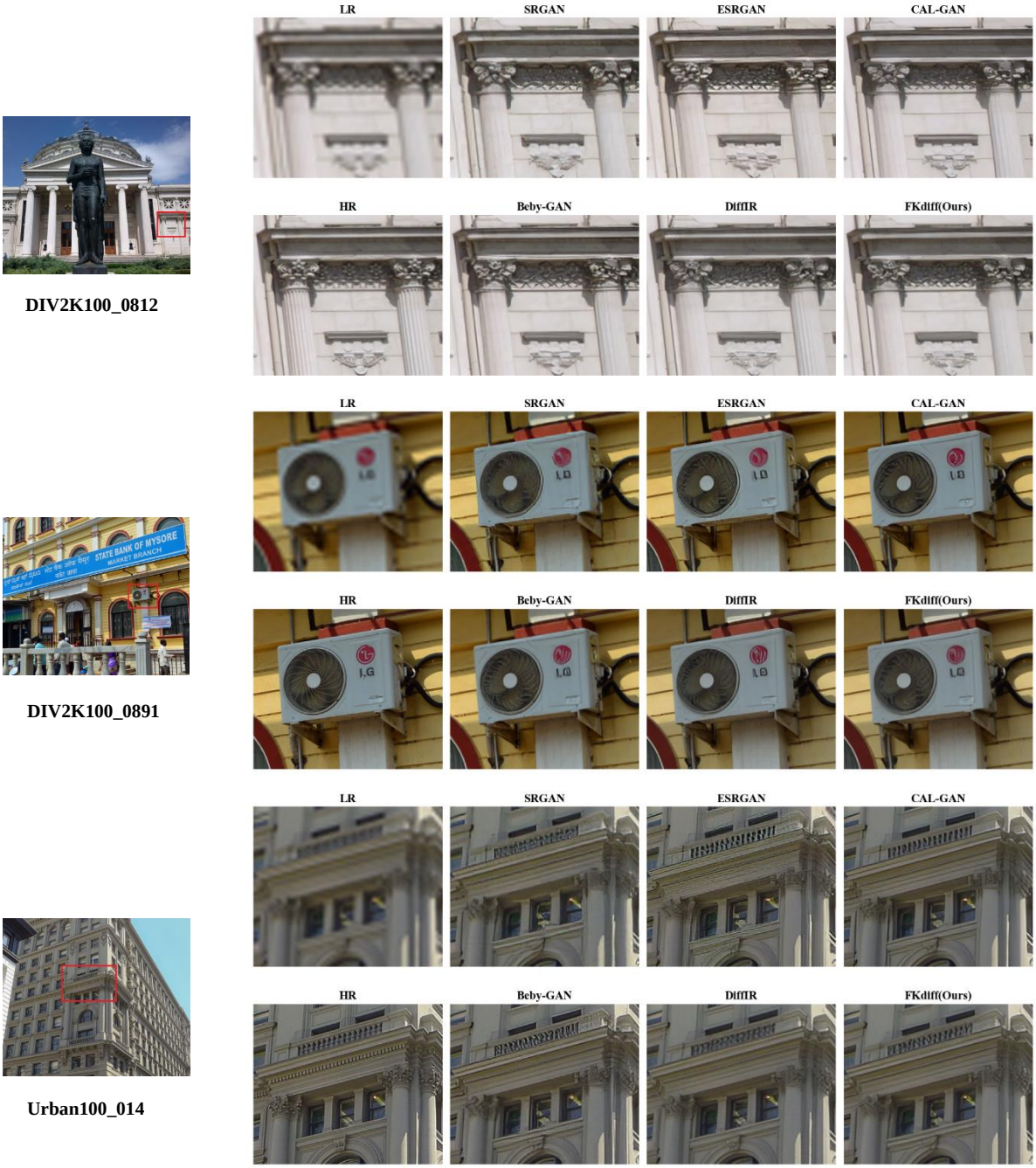

Figure 6. A comparison of visual results for 4x upscaling using different methods on DIV2K and Urban100. Images are zoomed in to better display the details.

Table 2. Quantitative evaluations at an image size of 256×256 using a single NVIDIA GeForce RTX 3090 GPU.

Method	Params (M)	Multi-Adds (T)	Avg. Runtime (s)	GPU Memory (G)
LDM	168.64	-	-	Out of memory
SinSR	118.59	-	1.58	3.42
DiffIR	26.57	0.70	0.1981	4.86
FKdiff	13.76	0.53	0.2106	2.79
RRDB	16.69	1.18	0.1507	0.83

4.3. Analysis

From Table 1, it is evident that our proposed model demonstrates several significant advantages over existing methods in single image super-resolution tasks. On the Manga109 dataset, our model achieved an LPIPS score of 0.0555, substantially better than ESRGAN's 0.0649 and CAL-GAN's 0.0699. Additionally, our model's PSNR and SSIM scores on certain datasets are comparable to those of DiffIR. For example, on the DIV2K100 dataset, our model achieved a PSNR of 29.05 vs. 29.11 and an SSIM of 0.7955 vs. 0.7958, indicating competitive performance. This slight difference underscores the effectiveness of our knowledge distillation method, showcasing our approach's competitiveness in high-quality image reconstruction and its ability to rival state-of-the-art diffusion models. Our model's outstanding performance across all evaluated datasets highlights its robustness and versatility.

We also evaluated the efficiency of our model, as detailed in Table 2. For different methods on 256×256 image sizes using an NVIDIA GeForce RTX 3090 GPU, we quantified the parameters (Params), computational cost (Multi-Adds), average runtime (Avg. runtime), and GPU memory consumption (GPU Memory). FKdiff has the fewest parameters at 13.76M. Its computational cost is 0.53T, significantly lower than DiffIR's 0.70T and RRDB's 1.18T. In terms of GPU memory consumption, FKdiff's 2.79G is far below DiffIR's 4.86G, using only 57% of its computational resources. While FKdiff's average runtime is 0.2106 seconds, this slightly longer runtime is acceptable considering its lower computational cost and GPU memory usage, especially in terms of resource savings.

In summary, our model excels in providing high perceptual quality and maintaining competitiveness across various datasets. The effectiveness of the knowledge distillation technique is evident, enabling our model to achieve results comparable to or even surpassing other leading methods in perceptual quality, including both diffusion-based and GAN-based methods.

5. ABLATION STUDY

We employed a hexagonal knowledge distillation architecture to distill knowledge into the target network. To demonstrate the efficiency of this architecture, we introduced $\text{FASRer}_{\text{XL}}$ as an intermediate network between FASRer_{L} and the teacher network, with the Transformer Block configuration set to [12, 2, 2, 4]. Notably, adding $\text{FASRer}_{\text{XL}}$ increased the training time by 48 hours, and the total number of training epochs was only one-third of the training time without $\text{FASRer}_{\text{XL}}$. Despite this, the impact on knowledge learning for the smaller network was limited.

Additionally, we evaluated the performance of the architecture when employing online distillation. Allowing the teacher network's parameters to remain unfrozen resulted in poorer outcomes. Specifically, the loss remained high throughout the training process without decreasing, and the generated images were excessively blurry, as shown in Figure 7(a). We concluded that the teacher network was being negatively impacted by the student network's interference, leading to deteriorating performance. Therefore, we ultimately decided to lock the teacher network's parameters.

We also validated the effectiveness of the Frequency Discriminative Distiller (FDD). Compared to the vanilla method, FDD allowed us to recover more details, as shown in Figure 7(b). The experimental results supporting these findings are summarized in Table 3, where "w/o" denotes removing and "PL" and "HA" represent "Parameter Lock" and "Hexagonal Architecture," respectively.

Table 3. Quantitative evaluations of each component in the proposed method on the DIV2K100 dataset.

	Parameter Lock	Hexagonal Architecture	FDD	PSNR/SSIM/LPIPS
w/o PL	×	✓	✓	26.35/0.7394/0.1647
w/o HA	✓	×	✓	28.95/0.7877/0.1031
w/o FDD	✓	✓	×	28.27/0.7785/0.1256
PL + HA + FDD	✓	✓	✓	29.05/0.7955/0.1036

(a)

(b)

Figure 7. In (a), the left image is the high-resolution (HR) image generated by FKdiff without parameter locking, while the right image is the version with parameter locking. In (b), the left image is the HR image produced by the vanilla distillation method, and the right image is generated using the Frequency-Domain Distillation (FDD) approach.

6. CONCLUSION

This paper introduces FKdiff, an innovative diffusion model designed for high-quality single-image super-resolution. FKdiff uniquely integrates Fourier frequency domain transformation and knowledge distillation, distinguishing it from existing methods. By utilizing low-resolution images as prior information and operating in the frequency domain, FKdiff achieves enhanced computational efficiency while preserving high-frequency details. The proposed Progressive Hexagonal Knowledge Distillation (PHexKD) approach ensures lightweight model deployment without compromising performance. Experimental results demonstrate that FKdiff outperforms existing methods in both efficiency and effectiveness, maintaining a high level of image generation quality.

REFERENCES

[1] J. Ho, A. Jain, and P.J.A.i.n.i.p.s. Abbeel, Denoising diffusion probabilistic models, 33 (2020), 6840-6851.

[2] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil, Model compression, in: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006, pp. 535-541.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684-10695.

[4] B.B. Moser, S. Frolov, F. Raue, S. Palacio, and A. Dengel, Dwa: Differential wavelet amplifier for image super-resolution, in: International Conference on Artificial Neural Networks, Springer, 2023, pp. 232-243.

[5] F. Guth, S. Coste, V. De Bortoli, and S.J.A.i.N.I.P.S. Mallat, Wavelet score-based generative modeling, 35 (2022), 478-491.

[6] S. Shang, Z. Shan, G. Liu, and J.J.a.p.a. Zhang, Resdiff: Combining cnn and diffusion model for image super-resolution, (2023).

[7] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R.J.a.p.a. Rombach, Sdxl: Improving latent diffusion models for high-resolution image synthesis, (2023).

[8] B. Xia, Y. Zhang, S. Wang, Y. Wang, X. Wu, Y. Tian, W. Yang, and L. Van Gool, Diffir: Efficient diffusion model for image restoration, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 13095-13105.

[9] Y. Wang, W. Yang, X. Chen, Y. Wang, L. Guo, L.-P. Chau, Z. Liu, Y. Qiao, A.C. Kot, and B. Wen, SinSR: diffusion-based image super-resolution in a single step, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 25796-25805.

[10] S. Zagoruyko and N.J.a.p.a. Komodakis, Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer, (2016).

[11] L. Kong, J. Dong, J. Ge, M. Li, and J. Pan, Efficient frequency domain-based transformers for high-quality image deblurring, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 5886-5895.

[12] G. Hinton, O. Vinyals, and J.J.a.p.a. Dean, Distilling the knowledge in a neural network, (2015).

[13] N. Passalis, M. Tzelepi, and A. Tefas, Heterogeneous knowledge distillation using information flow modeling, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2339-2348.

[14] C. Fang, Q. Wang, L. Cheng, Z. Gao, C. Pan, Z. Cao, Z. Zheng, and D.J.I.T.o.M.I. Zhang, Reliable mutual distillation for medical image segmentation under imperfect annotations, (2023).

[15] Q. Guo, X. Wang, Y. Wu, Z. Yu, D. Liang, X. Hu, and P. Luo, Online knowledge distillation via collaborative learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11020-11029.

[16] Z. He, T. Dai, J. Lu, Y. Jiang, and S.-T. Xia, Fakd: Feature-affinity based knowledge distillation for efficient image super-resolution, in: 2020 IEEE International Conference on Image Processing (ICIP), IEEE, 2020, pp. 518-522.

[17] W. Lee, J. Lee, D. Kim, and B. Ham, Learning with privileged information for efficient image super-resolution, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16, Springer, 2020, pp. 465-482.

[18] C. Pham, V.-A. Nguyen, T. Le, D. Phung, G. Carneiro, and T.-T. Do, Frequency attention for knowledge distillation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 2277-2286.

[19] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H.U.J.a.p.a. Li, A general u-shaped transformer for image restoration. arXiv 2021.

[20] D. Fuoli, L. Van Gool, and R. Timofte, Fourier space losses for efficient perceptual image super-resolution, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2360-2369.

[21] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, Enhanced deep residual networks for single image super-resolution, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2017, pp. 136-144.

[22] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, and L. Zhang, Ntire 2017 challenge on single image super-resolution: Methods and results, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2017, pp. 114-125.

[23] R. Zeyde, M. Elad, and M. Protter, On single image scale-up using sparse-representations, in: Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7, Springer, 2012, pp. 711-730.

[24] J.-B. Huang, A. Singh, and N. Ahuja, Single image super-resolution from transformed self-exemplars, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 5197-5206.

[25] D. Martin, C. Fowlkes, D. Tal, and J. Malik, A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics, in: Proceedings eighth IEEE international conference on computer vision. ICCV 2001, IEEE, 2001, pp. 416-423.

[26] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki, K.J.M.t. Aizawa, and applications, Sketch-based manga retrieval using manga109 dataset, 76 (2017), 21811-21838.

[27] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, and Z. Wang, Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4681-4690.

[28] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, Esrgan: Enhanced super-resolution generative adversarial networks, in: Proceedings of the European conference on computer vision (ECCV) workshops, 2018, pp. 0-0.

[29] W. Li, K. Zhou, L. Qi, L. Lu, and J. Lu, Best-buddy gans for highly detailed image super-resolution, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2022, pp. 1412-1420.

[30] J. Park, S. Son, and K.M. Lee, Content-aware local gan for photo-realistic super-resolution, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 10585-10594.

[31] H. Li, Y. Yang, M. Chang, S. Chen, H. Feng, Z. Xu, Q. Li, and Y.J.N. Chen, Srdiff: Single image super-resolution with diffusion probabilistic models, 479 (2022), 47-59.

[32] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, and O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 586-595.