Degradation-Aware Frequency-Separated Transformer for Blind Super-Resolution

Hanli Zhao
Wenzhou University, Wenzhou, China
hanlizhao@wzu.edu.cn

Binhao Wang
Wenzhou University, Wenzhou, China
binhaowangfun@163.com

Wanglong Lu✉
Wenzhou University, Wenzhou, China
wanglongl@mun.ca

Juncong Lin
Xiamen University, Xiamen, China
jclin@xmu.edu.cn

Abstract

Blind image super-resolution involves reconstructing high-resolution images from low-resolution inputs with various unknown degradations. It is a challenging task due to the limited information available from the degraded images. While existing methods have achieved impressive results, they often overlook high-frequency or low-frequency features, reducing their effectiveness. To solve this problem, we propose a frequency-separated Transformer framework with degradation-aware learning for blind super-resolution. We first introduce a multi-patch contrastive learning approach to implicitly learn discriminative degradation representations. To fully utilize degradation representations as guidance information, a frequency-separated self-attention mechanism is introduced to extract global structural and local detail features separately. Our degradation-aware frequency-separated Transformer progressively restores high-quality images using successive frequency-separated self-attention blocks. Extensive experiments demonstrate that our approach outperforms state-of-the-art methods on four benchmark blind super-resolution datasets, while also achieving lower GPU memory usage during training and faster inference speed.

Keywords: ***Image super-resolution, Blind super-resolution, Contrastive learning, Degradation representation, Transformers***

1. Introduction

Blind image super-resolution (BSR) aims to reconstruct high-resolution (HR) images from low-resolution (LR) input images with unknown degradation. This challenging task is a crucial subset of the broader field of image restoration [24] in computer vision, including image denoising [42], dehazing [37], and various forms of super-resolution [3], all sharing the common goal of reconstructing high-quality images from degraded inputs.

In the realm of super-resolution, numerous outstanding approaches have been developed, leveraging both convolutional neural networks (CNN) [6, 52] and Transformer architectures [19, 4]. However, these methods are primarily non-blind and operate under the assumption of a specific, predefined degradation type, such as bicubic downsampling. When confronted with degradations that deviate from this assumption, models trained on fixed degradation conditions, often struggle to deliver satisfactory outcomes.

To achieve high-quality BSR, two widely recognized challenges dominate recent literature. The first one is accurately predicting the degradation type from a given image, which is crucial for providing more effective guidance during the subsequent super-resolution process. However, existing methods typically rely on limited perspectives to learn implicit degradation representation, potentially restricting their ability to capture the degradation representations [36, 17]. The exploration of multi-patch approaches for BSR remains relatively underexplored in current research. The second challenge is efficiently utilizing the predicted degradation representations to extract deep-level features. While the goal is to fully leverage the obtained representations to handle complex degradations, CNN-based methods [36, 17] often lack the large receptive field necessary for capturing long-range dependencies in features. On the other hand, Transformer-based methods [22, 30] are better at modeling global context but often compromise by using moderately sized windows to balance the computational demands. Some methods further utilize high- and low-frequency features [28] to boost the extraction of global and local information. However, research for the modeling of high- and low-frequency features based on degradation guidance remains limited.

In this paper, we explore the ways to address these chal-

lenges. We first leverage multiple patches from a single image, to gain diverse perspectives, leading to more robust representations and reducing overfitting risks [2, 53]. Second, recognizing that image features include high frequencies for fine details and low frequencies for global structures [28], we propose a conditional dual-path self-attention mechanism guided by degradation representations. The high-frequency path captures local details with small windows, while the low-frequency path models global interactions on compact feature maps. This approach boosts efficiency and both local and global feature modeling.

To this end, we propose a novel degradation-aware frequency-separated Transformer (DAFST) for blind super-resolution. First, we introduce a multi-patch contrastive learning approach to learn discriminative degradation representations implicitly. These representations serve as degradation information to guide the super-resolution process in our proposed frequency-separated attention blocks. Each block utilizes a frequency-separated self-attention mechanism to extract global structural and local detail features separately. The frequency-specific features are then fused in each block, progressively restoring high-quality images through successive blocks. We have extensively compared our method against the SOTA models and conducted comprehensive experiments to demonstrate its superiority. The contributions of this paper are as follows:

- We introduce a novel degradation-aware frequency-separated Transformer (DAFST) that effectively leverages implicit degradation representations to enhance feature extraction for high-quality blind super-resolution.
- We propose a degradation-guided frequency-separated attention mechanism that decouples global and local detail modeling, enhancing both modeling capabilities and inference efficiency.
- We design a multi-patch contrastive learning approach to capture degradation representations from diverse perspectives, leading to more robust features.
- Our method achieves state-of-the-art performance across multiple datasets and degradation types while maintaining faster inference speeds.

2. Related work

2.1. Non-blind image super-resolution

Image super-resolution is a fundamental problem in computer vision, aiming to reconstruct an HR image from an LR image. Early methods for image super-resolution predominantly relied on CNN, extensively discussed in the literature [45]. CNN-based SR methods, such as studies like [7, 48, 18, 10], gained prominence due to their excellent local inductive bias. SRCNN [8], a seminal work in image super-resolution, employed a simple three-layer CNN to learn the LR-HR mapping for image super-resolution, catalyzing numerous subsequent advancements in the field.

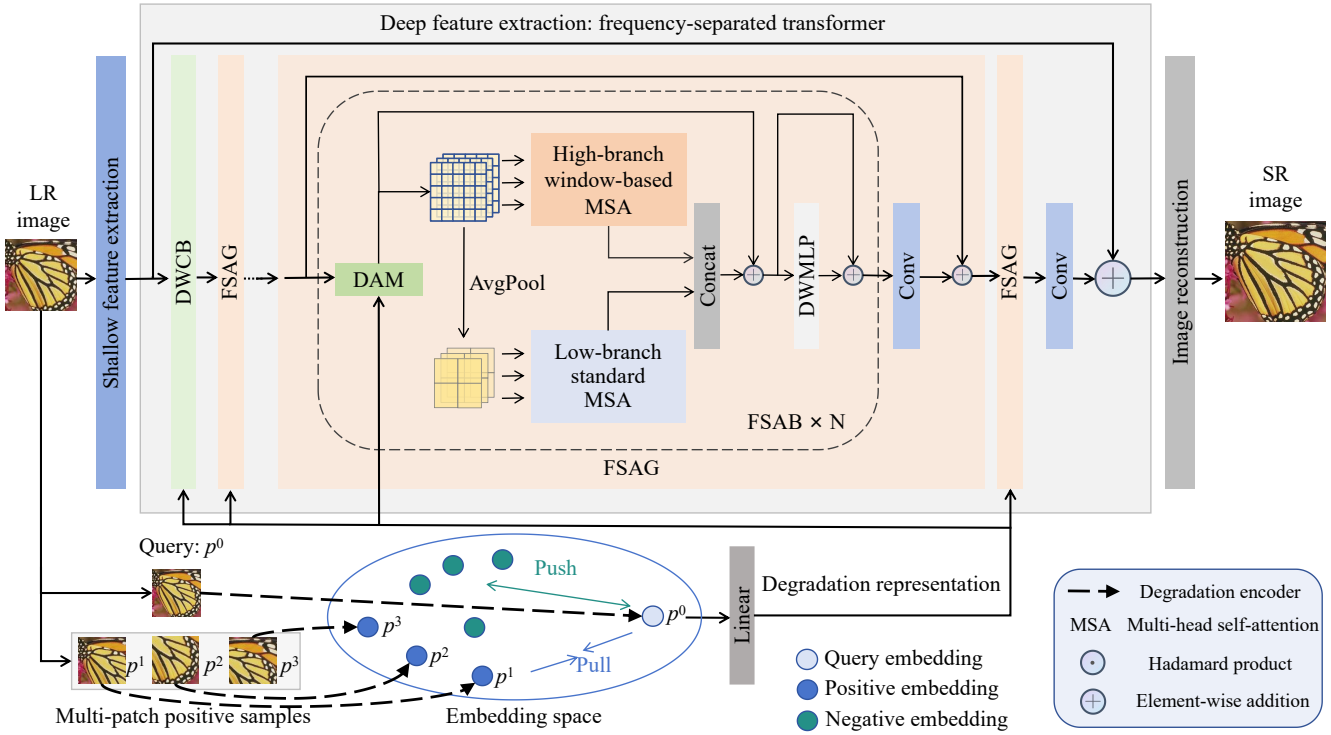
Typically, super-resolution methods comprise three primary modules: shallow feature extraction, deep feature extraction, and super-resolution reconstruction modules. In recent years, significant enhancements have been made to the deep feature extraction and super-resolution reconstruction modules, with network processing strategies such as upsampling [9], residual learning [15], and sub-pixel convolutional upsampling [31] becoming standard paradigms for constructing super-resolution networks. With advancements in methods such as RCAN [52], SAN [6], and SwinIR [19], the performance of non-blind image super-resolution on fixed degradation types, like bicubic downsampling, has reached a plateau. Recent works, such as CAMixer [39], propose sampling convolution or visual attention based on the complexity of image regions, and MambaIR [12] utilizes 2D-selective-scan to replace traditional attention mechanisms. While these methods have achieved remarkable results, non-blind approaches often struggle to generalize effectively to unknown degradations beyond their predefined scope.

2.2. Blind image super-resolution

Networks designed for fixed bicubic downsampling often suffer significant performance drops when faced with real-world degradations. Several blind super-resolution methods were proposed to tackle this challenge. Currently, three primary methods are employed for obtaining degradation: (1) Adapted non-blind super-resolution, which assumes known degradation and uses it as a prior. (2) Blind super-resolution based on explicit kernel estimation. (3) Blind super-resolution based on implicit degradation representation.

Adapted non-blind super-resolution methods. These methods explicitly incorporate blur kernels as additional inputs to guide the SR process. For instance, SRMD [50] and UDVD [43] introduced degradation kernels as inputs to handle varying degradation conditions, while DPSR [51] and USRNet [47] leveraged variable splitting techniques to optimize energy functions, treating blur kernels as independently optimized terms. However, these methods are highly dependent on the accurate input from the blur kernel, limiting their generalization to a broader degradation space. When faced with unseen degradations or without precise kernel information, their performance significantly degrades.

Explicit kernel estimation methods. Other methods aim to guide the network by estimating blur kernels directly from the LR image. For example, Liang et al. [20]



(a) The architecture of our degradation-aware frequency-separated Transformer (DAFST).

Figure 1: Our DAFST integrates degradation representations with image features through multi-patch contrastive learning, the images are encoded into embeddings with the degradation encoder, aiming to maximize the similarity between query embeddings and positive embeddings while minimizing the similarity with negative embeddings. Guided by these representations, the frequency-separated self-attention blocks extract global and local features, which are then merged using depthwise convolutions and reconstructed via the image reconstruction module.

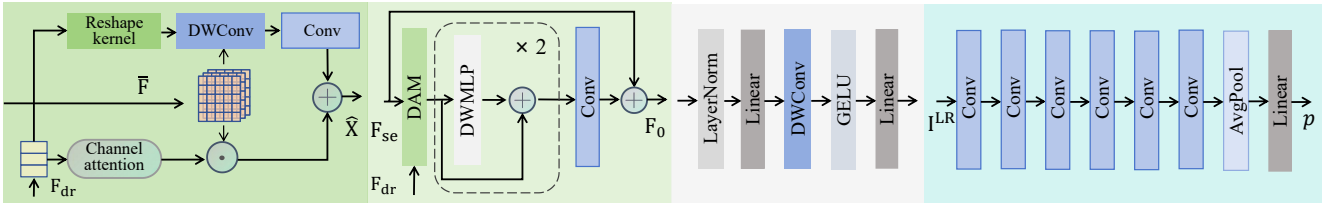
proposed using multiple mini super-resolution network experts to estimate diverse degradations. Gu [11] introduced an iterative kernel correction (IKC) method that refines the estimated degradation based on intermediate SR results. ZSSR [32] leverages internal image recurrence, utilizing repetitive structures within the LR image for iterative optimization. However, these methods often incur high computational costs due to multiple iterations of kernel estimation and correction during testing. To address this, DCLS [25] and MZSR [33] build upon IKC and ZSSR, respectively, reducing the number of iterative optimizations. Similarly, Luo et al. [26] developed a deep alternating network (DAN) to iteratively estimate degradation and restore SR images. Despite these advancements, these methods remain highly sensitive to the accuracy of kernel estimation, and inaccurate estimations can lead to suboptimal results, limiting their robustness in real-world scenarios.

Implicit degradation representation methods. Recently, some studies have shown that implicit degradation representation is better suited to handle complex degradation scenarios. DASR [36] pioneered the integration of contrastive learning to guide deep feature extraction. Wei et al. proposed [40] unsupervised domain gap-aware training networks. Additionally, DSSR [17] introduced a detail structure modulation module to enhance details cyclically. KDSR [41] introduced knowledge distillation to enable a student model to learn degradation representations from another teacher model. DSAT [22] introduced residual Swin-Transformer blocks [23] to address the limited receptive field issue of CNN, setting a new benchmark in BSR. This paper proposes a novel degradation-aware frequency-separated Transformer with multi-patch contrastive learning to improve implicit degradation representation learning.

3. Methods

3.1. Overview

Fig. 1 (a) shows the overall pipeline of our degradation-aware frequency-separated Transformer (DAFST) for blind super-resolution. Our DAFST takes a low-resolution image $\mathbf{I}^{\text{LR}}$ as input and produces a super-resolution image $\mathbf{I}^{\text{SR}}$, which is obtained as $\mathbf{I}^{\text{SR}} = \text{DAFST}(\mathbf{I}^{\text{LR}})$. The DAFST



(a) DAM (b) DWCB (c) DWMLP (d) Degradation encoder

Figure 2: The details of our degradation-aware modulation (DAM), depthwise convolution block (DWCB), DWMLP, and degradation encoder. The DWConv means depthwise convolution.

network consists of four key components: degradation encoder, shallow feature extraction, deep feature extraction, and image reconstruction modules.

Degradation encoder. We apply a degradation encoder and a linear projection to learn distinctive degradation representations. The learned degradation representations guide the subsequent modules in extracting deep features. It can be represented as:

$$\mathbf{F}_{dr} = \text{Linear}(\text{Encoder}(\mathbf{I}^{\text{LR}})), \tag{1}$$

where $\mathbf{F}_{dr} \in \mathbb{R}^{256}$ signifies the degradation representations; $\text{Encoder}(\cdot)$ denotes the degradation encoder. We employ multi-patch contrastive learning to train our degradation encoder.

Shallow feature extraction module. To extract shallow features, a simple 3×3 convolution layer is used as the shallow feature extraction module $\text{H}_{\text{SE}}(\cdot)$. It can be represented as:

$$\mathbf{F}_{se} = \text{H}_{\text{SE}}(\mathbf{I}^{\text{LR}}), \tag{2}$$

where $\mathbf{I}^{\text{LR}} \in \mathbb{R}^{H\times W\times 3}$; $\mathbf{F}_{se} \in \mathbb{R}^{H\times W\times C}$ denotes the extracted shallow features.

Frequency-separated Transformer. The deep feature extraction involves a novel frequency-separated Transformer for BSR. It consists of a depthwise convolution block (DWCB) and groups of frequency-separated self-attention blocks (FSAB). Each block aims to extract deep features for reconstruction. The deep feature extraction module receives the feature maps $\mathbf{F}_{se}$ and degradation representations $\mathbf{F}_{dr}$, and output deep features $\mathbf{F}_{de} \in \mathbb{R}^{H\times W\times C}$, which is described as follows:

$$\mathbf{F}_{de} = \text{H}_{\text{DE}}(\mathbf{F}_{se}, \mathbf{F}_{dr}), \tag{3}$$

where $\text{H}_{\text{DE}}(\cdot)$ represents the deep extraction module.

Image reconstruction module. The image reconstruction module $\text{H}_{\text{REC}}(\cdot)$ consists of a convolutional layer and a sub-pixel convolutional upsampling layer [9]. We reconstruct high-quality images $\mathbf{I}^{\text{SR}}$ by fusing shallow features and deep features as illustrated below:

$$\mathbf{I}^{\text{SR}} = \text{H}_{\text{REC}}(\mathbf{F}_{se} + \mathbf{F}_{de}). \tag{4}$$

3.2. Frequency-separated Transformer

Given the shallow feature maps $\mathbf{F}_{se}$, our frequency-separated Transformer $\text{H}_{\text{DE}}(\cdot)$ consists of two stages. The input features first go through a depthwise convolution block: $\mathbf{F}_0 = \text{DWCB}(\mathbf{F}_{se}, \mathbf{F}_{dr})$. Then the intermediate features are continuously extracted through frequency-separated self-attention groups, which are expressed as:

$$\mathbf{F}_s = \text{H}_{\text{FSAG}_s}(\mathbf{F}_{s-1}, \mathbf{F}_{dr}), \quad s = 1, 2, \ldots, S, \tag{5}$$

where $\text{DWCB}(\cdot)$ is the depthwise convolution block to extract shallow features; The $\text{H}_{\text{FSAG}_s}(\cdot)$ represents s-th frequency-separated self-attention group. Our frequency-separated Transformer contains S FSAG groups. After S groups feature extraction, we use a 3×3 convolution to get the output deep features $\mathbf{F}_{de} = \text{Conv}_{3\times3}(\mathbf{F}_S)$.

Degradation-aware modulation (DAM). As shown in Fig. 2 (a), the degradation-aware modulation (DAM) mechanism is designed to integrate degradation representations during the super-resolution process. We use channel attention to obtain channel weights $\mathbf{F}_{co}$ from the degradation representations $\mathbf{F}_{dr}$ and to modulate convolutional layers for guidance-based feature extraction. Given the $\mathbf{F}_{dr}$, the generation of channel coefficients $\mathbf{F}_{co} \in \mathbb{R}^{H\times W\times C}$ can be expressed as:

$$\begin{aligned}\mathbf{F}_{co} &= \text{ChannelAttention}(\mathbf{F}_{dr}) \\ &= \text{Sigmoid}\left(\text{Conv}_{1\times1}\left(\text{ReLU}\left(\text{Conv}_{1\times1}(\mathbf{F}_{dr})\right)\right)\right),\end{aligned} \tag{6}$$

where $\text{Sigmoid}(\cdot)$ is the sigmoid activation; $\text{Conv}_{1\times1}(\cdot)$ is the 1×1 convolution layer; $\text{ReLU}(\cdot)$ is the LeakyReLU activation function with the 0.1 negative slope. Then we can get the degradation modulated depthwise convolution kernel:

$$\text{DWConv}_{\mathbf{F}_{dr}} = \text{GeneratedKernel}(\mathbf{F}_{dr}) = \text{Reshape}(\mathbf{F}_{dr}), \tag{7}$$

We reshape $\mathbf{F}_{dr}$ as a convolution kernel $\in \mathbb{R}^{C\times1\times3\times3}$ denoted as $\text{DWConv}_{\mathbf{F}_{dr}}(\cdot)$ and get $\hat{\mathbf{X}} \in \mathbb{R}^{H\times W\times C}$ from the input feature tensor $\bar{\mathbf{F}}$. As shown in Fig. 2 (a), we ultimately obtain $\hat{\mathbf{X}}$ through the following operations:

$$\hat{\mathbf{X}} = \text{Conv}_{1\times1}(\text{DWConv}_{\mathbf{F}_{\text{dr}}}(\text{ReLU}(\bar{\mathbf{F}}))) + (\bar{\mathbf{F}} \odot \mathbf{F}_{co}), \tag{8}$$

Algorithm 1 Frequency-separated self-attention block

Input: Input feature tensor $\bar{\mathbf{F}}$, degradation representations $\mathbf{F}_{dr}$
Parameters: Number of total attention heads O, number of high-frequency branch attention heads O_1, number of low-frequency branch attention heads O_2, and channel dimension per head C_o
Degradation-aware modulation
$\mathbf{F}_{co} = \text{ChannelAttention}(\mathbf{F}_{dr})$
$\text{DWConv}_{\mathbf{F}_{dr}} = \text{GeneratedKernel}(\mathbf{F}_{dr})$
$\hat{\mathbf{X}} = \text{Conv}_{1\times1}(\text{DWConv}_{\mathbf{F}_{dr}}(\text{ReLU}(\bar{\mathbf{F}}))) + (\bar{\mathbf{F}} \odot \mathbf{F}_{co})$
Frequency-separated self-attention
$\bar{\mathbf{X}} = \text{Flatten}(\hat{\mathbf{X}})$ # $\bar{\mathbf{X}} \in \mathbb{R}^{L\times C}$
$\mathbf{X}^h, \mathbf{X}^l = \text{Divide}(\bar{\mathbf{X}})$ # $\mathbf{X}^h \in \mathbb{R}^{O_1\times L\times C_o}$, $\mathbf{X}^l \in \mathbb{R}^{O_2\times L\times C_o}$
Combine high and low frequency attention
$\mathbf{Y}' = [\text{H-MSA}(\text{LN}(\mathbf{X}^h)); \text{L-MSA}(\text{LN}(\text{AvgPool}(\mathbf{X}^l)))] + \bar{\mathbf{X}}$
$\mathbf{Y}'' = \text{DWMLP}(\mathbf{Y}') + \mathbf{Y}'$
return $\mathbf{Y}''$

where $\hat{\mathbf{X}}$ is the intermediate feature obtained by fusing the input and degradation representations and $\odot$ denotes Hadamard product with broad-casting technique.

Depthwise convolution block (DWCB). As shown in Fig. 2 (b), our depthwise convolution block (DWCB) consists of a DAM and two depthwise multilayer perception (DWMLP) blocks. As shown in Fig. 2 (c), each DWMLP block contains two linear projections and a 3×3 depthwise convolution (DWConv) layer. It extracts shallow features and implicitly learns position encoding for the subsequent feature extraction.

Frequency-separated self-attention group (FSAG). Each FSAG takes degradation representations $\mathbf{F}_{dr}$ and aggregates input image features to extract deeper features. Each FSAG consists of N FSAB blocks; each FSAB also leverages $\mathbf{F}_{dr}$ to enhance feature aggregation from the input images. Given the input features $\mathbf{Y}_0 = \mathbf{F}_{s-1}$, for the s-th group $\text{H}_{\text{FSAG}_s}(\cdot)$ with N frequency-separated attention blocks, the deep features $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_N$ are extracted sequentially. This process in each group, defined as $\mathbf{F}_s = \text{H}_{\text{FSAG}_s}(\mathbf{F}_{s-1}, \mathbf{F}_{dr})$, and $\mathbf{F}_s = \mathbf{Y}_N$. It can be described as follows:

$$\begin{aligned} \mathbf{Y}_n &= \text{H}_{\text{FSAB}_n}(\mathbf{Y}_{n-1}, \mathbf{F}_{dr}), \quad n = 1, 2, \ldots, N-1, \\ \mathbf{Y}_N &= \text{Conv}_{3\times3}(\text{H}_{\text{FSAB}_N}(\mathbf{Y}_{N-1}, \mathbf{F}_{dr})) + \mathbf{Y}_0, \end{aligned} \tag{9}$$

where $\text{H}_{\text{FSAB}_n}(\cdot)$ denotes the n-th frequency-separated self-attention block within the s-th FSAB group.

Frequency-separated self-attention block (FSAB). The FSAB primarily comprises two key components, the DAM and a frequency-separated self-attention mechanism. This design enables the block to effectively process and integrate degradation information and image features across different frequency bands. More details are shown in Algorithm 1.

Frequency-separated self-attention. Our frequency-separated self-attention mechanism consists of high-frequency and low-frequency multi-head self-attention (MSA) modules. The high-frequency attention focuses on capturing fine details, while the low-frequency attention extracts global information, such as structural elements, to enhance super-resolution performance. The input tensor $\hat{\mathbf{X}} \in \mathbb{R}^{H\times W\times C}$ is initially flattened into $\bar{\mathbf{X}} \in \mathbb{R}^{L\times C}$, where $L = H \times W$, which is denoted as $\bar{\mathbf{X}} = \text{Flatten}(\hat{\mathbf{X}})$. Then, the tensor $\bar{\mathbf{X}}$ is divided along with the channel dimension into high-frequency $\mathbf{X}^h \in \mathbb{R}^{O_1\times L\times C_o}$ and low-frequency $\mathbf{X}^l \in \mathbb{R}^{O_2\times L\times C_o}$ tensors, respectively. $O = O_1 + O_2$ represents the total number of attention heads, comprising O_1 high-frequency and O_2 low-frequency attention heads. $C_o = C/O$ corresponds to the channel number for each head. For the each frequency-separated self-attention feature extraction, denoted as $\mathbf{Y}_n = \text{H}_{\text{FSAB}_n}(\mathbf{Y}_{n-1}, \mathbf{F}_{dr})$, we have $\mathbf{Y}_n = \mathbf{Y}''$. Our frequency-separated self-attention can be expressed as:

$$\mathbf{Y}' = [\text{H-MSA}(\text{LN}(\mathbf{X}^h)); \text{L-MSA}(\text{LN}(\text{AvgPool}(\mathbf{X}^l)))] + \bar{\mathbf{X}}, \tag{10}$$

$$\mathbf{Y}'' = \text{DWMLP}(\mathbf{Y}') + \mathbf{Y}', \tag{11}$$

where $\text{H-MSA}(\cdot)$ and $\text{L-MSA}(\cdot)$ are features extracted from our high-frequency and low-frequency multi-head attention modules. The $[\cdot]$ denotes the concatenation operation. $\text{AvgPool}(\cdot)$ is the average pooling operation. We concatenate the outputs from the dual branches, and the final multi-head attention can be calculated. A LayerNorm (LN) layer is added before the multi-head self-attention, and residual connections are utilized in each block.

Here, we design a high-frequency self-attention module to extract high-frequency features, by adopting a window-based attention mechanism. The input $\mathbf{X}^h$ is partitioned into non-overlapping windows and reshaped to $O_1 \times \frac{L}{M^2} \times M^2 \times C_o$, where M represents the local window size. We set $M = 4$, using small local windows of 4×4 for self-attention to capture high-frequency

fine-grained features instead of larger window sizes such as 8×8 or 16×16. Moreover, we do not use techniques like shift window [19] or multi-scale windows [44], which saves considerable computational complexity and makes our approach hardware-friendly. The high-frequency MSA branch can be described as: $\text{H-MSA}(\mathbf{X}^h) = [\text{SA}_1(\mathbf{X}_1^h), \ldots, \text{SA}_o(\mathbf{X}_o^h), \ldots, \text{SA}_{O_1}(\mathbf{X}_{O_1}^h)]\mathbf{W}_{O_1}$. The o indicates the head index and $\mathbf{X}_o^h \in \mathbb{R}^{\frac{L}{M^2} \times M^2 \times C_o}$ is each single head. The projection matrix $\mathbf{W}_{O_1} \in \mathbb{R}^{(O_1 \times C_o) \times (O_1 \times C_o)}$ and $[\cdot]$ concatenates the outputs from the O_1 attention heads. The $\text{SA}_o(\cdot)$ is the self-attention module.

Here, we design a low-frequency self-attention to capture global low-frequency features. Since it is unfeasible to perform global attention directly in the entire pixel space due to the high computational cost, we employ an average pooling to downsample the $\mathbf{X}^l$ into $\mathbf{X}' \in \mathbb{R}^{O_2 \times (L/D^2) \times C_o}$, which can be defined as: $\mathbf{X}' = \text{AvgPool}(\mathbf{X}^l)$, where D means the pooling kernel size and we set $D = 4$ in this paper. $\text{AvgPool}(\cdot)$ is used to encode inputs into a lower-dimensional latent space for extracting low-frequency global information. Then, we use a standard multi-head self-attention mechanism in the latent space to capture rich low-frequency information from the feature maps. We then have $\text{L-MSA}(\mathbf{X}') = [\text{SA}_1(\mathbf{X}'_1), \ldots, \text{SA}_o(\mathbf{X}'_o), \ldots, \text{SA}_{O_2}(\mathbf{X}'_{O_2})]\mathbf{W}_{O_2}$ for the low-frequency MSA branch. $\mathbf{W}_{O_2} \in \mathbb{R}^{(O_2 \times C_o) \times (D^2 \times O_2 \times C_o)}$ is a projection matrix.

Now, we describe the details of the self-attention for o-th head $\text{SA}_o(\mathbf{X})$. Given the input feature maps $\mathbf{X} \in \mathbb{R}^{L \times C_o}$, $\mathbf{X}$ is used to compute the query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ matrices, each with dimensions $L \times C_o$. These matrices are derived through distinct linear transformations applied to $\mathbf{X}$, which can be expressed as:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_q, \mathbf{K} = \mathbf{X}\mathbf{W}_k, \mathbf{V} = \mathbf{X}\mathbf{W}_v, \tag{12}$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{C_o \times C_o}$ are learnable parameters, and C_o is the number of hidden dimensions per head. Next, the output of a self-attention head is obtained by the Softmax activation function applied to the scaled dot product of the query and key:

$$\text{SA}_o(\mathbf{X}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{C_o}}\right)\mathbf{V}. \tag{13}$$

As depicted in Fig. 1 (a), the FSAB utilizes a low-frequency attention branch to capture the global dependencies inherent in the input image. This branch focuses on global attention and does not necessitate high-resolution feature maps. Conversely, the high-frequency attention branch is tailored to detect finely detailed local dependencies, utilizing a local attention mechanism on high-resolution feature maps to achieve this objective.

3.3. Loss functions

Multi-patch degradation representation learning. Contrastive learning has been used in unsupervised representation learning widely, such as MoCo [13], which aims to maximize mutual information within the representation embedding space by encouraging similar samples to be closer and dissimilar ones to be farther apart. To learn more robust features and reduce the risk of overfitting for single-view-based constructive learning, we propose to design a multiview scheme for contrastive learning approach to capture degradation representations from diverse perspectives.

Our degradation representation learning module consists of a contrastive learning encoder and an additional linear projection, which aims to capture degradation representations from LR images unsupervised implicitly. Specifically, a random image patch is cropped from an LR image as a query patch, other patches belonging to the same LR image are considered positive samples, and patches from different LR images are considered negative samples. Our encoder embeds the input into $\boldsymbol{p} \in \mathbb{R}^{256}$ following the main structure of MoCo [13], as shown in Fig. 2 (d).

To learn discriminative degradation representations, a large set of negative samples is essential [16]. Instead of relying on large batch sizes, our model maintains a queue of diverse samples to achieve content-invariant degradation representations while enabling concurrent training of degradation learning and feature extraction modules.

As shown in Fig. 1, we randomly select batch-size LR images to represent different degradations. For i-th sampled LR image $\mathbf{I}_i^{\text{LR}}$, it can be expressed as:

$$\boldsymbol{p}_i^t = \text{Encoder}(\hat{\mathbf{I}}_{i,t}^{\text{LR}}), t \in [0, 1, 2, \ldots, T], \tag{14}$$

where $\hat{\mathbf{I}}_{i,t}^{\text{LR}}$ is a randomly cropped image patch with different sizes (e.g., 32×32, 48×48, and 64×64) from the i-th image. For each image, we set $\boldsymbol{p}_i^0$ as a query embedding $\boldsymbol{q}_i \in \mathbb{R}^{256}$ and the others as T positive embeddings $\boldsymbol{p}_i^t \in \mathbb{R}^{256}, t \in [1, 2, \ldots, T]$. Conversely, any $\boldsymbol{p}_j, j \neq i$ belonging to other LR images is treated as a negative sample embedding.

Multi-patch contrastive loss. We employ multi-patch contrastive loss as the degradation loss to optimize our contrastive learning encoder, incorporating a temperature coefficient τ, which is defined by the following equation:

$$\mathcal{L}_{\text{degrade}} = -\frac{1}{T}\sum_{t=1}^{T}\log\frac{\exp(\boldsymbol{q}_i \cdot \boldsymbol{p}_i^t/\tau)}{\sum_{j=1}^{U}[\exp(\boldsymbol{q}_i \cdot \boldsymbol{q}_j/\tau) + \sum_{t=1}^{T}\exp(\boldsymbol{q}_i \cdot \boldsymbol{p}_j^t/\tau)]}, \tag{15}$$

where U represents the number of samples in the negative sample queue. $\boldsymbol{q}_j$ and $\boldsymbol{p}_j^t$ denote the negative embeddings from j-th sample in the queue. We set temperature coefficient $\tau = 0.07$ to control the sharpness of the Softmax function used in the computation of the loss function. The queue size U is set to 1024.

Table 1: Quantitative comparison on noise-free degradation and isotropic Gaussian kernels. The kernel widths (σ) are given for each column. The best and second-best results are marked in **bold** and underlined, respectively.

Method	Source	Scale	Set5			Set14			BSD100			Urban100		
			0.6	1.2	1.8	0.6	1.2	1.8	0.6	1.2	1.8	0.6	1.2	1.8
Bicubic			32.30	29.28	27.07	29.21	27.13	25.47	28.76	26.93	25.51	26.13	24.46	23.06
RCAN [52]	ECCV 2018		35.91	32.31	28.50	32.31	28.48	26.33	31.16	28.04	26.26	29.80	25.38	23.44
SRMD [50]	CVPR 2018		34.77	34.13	33.80	31.35	30.78	30.18	30.33	29.89	29.20	28.42	27.43	27.12
IKC [11]	CVPR 2019		37.35	37.26	33.94	33.36	32.97	30.31	31.97	31.79	29.57	31.37	30.53	27.15
SwinIR [19]	ICCV 2021		35.96	31.21	28.51	32.38	28.49	26.33	31.19	28.04	26.26	29.92	25.39	23.45
DASR [36]	CVPR 2021	$\times 2$	37.47	37.19	35.43	32.96	32.78	31.60	31.78	31.71	30.54	30.71	30.36	28.95
DAN [26]	Arxiv 2021		37.83	37.46	35.76	33.33	33.20	31.81	32.06	31.88	30.51	31.14	30.71	29.04
DSSR [17]	TMM 2022		37.94	37.60	**35.86**	33.40	33.29	**32.03**	32.15	32.06	30.88	31.42	31.15	**29.67**
HAT [5]	CVPR 2023		36.04	31.22	28.52	32.56	28.51	26.34	31.29	28.06	26.27	30.17	25.40	23.45
DSAT [22]	TMM 2024		38.06	37.59	35.65	33.55	33.34	31.88	32.21	32.10	30.88	31.50	31.03	29.40
DAFST (ours)			**38.09**	**37.65**	35.75	**33.78**	**33.57**	31.94	**32.25**	**32.11**	**30.89**	**31.76**	**31.18**	29.38
Method	Source	Scale	Set5			Set14			BSD100			Urban100		
			0.8	1.6	2.4	0.8	1.6	2.4	0.8	1.6	2.4	0.8	1.6	2.4
Bicubic			29.42	27.24	25.39	26.84	25.42	24.09	26.72	25.52	24.41	24.02	22.95	21.89
RCAN [52]	ECCV 2018		32.90	29.12	26.75	29.49	26.75	24.99	28.56	26.55	25.18	26.89	24.89	22.30
SRMD [50]	CVPR 2018		32.63	32.27	28.62	29.25	28.01	26.90	28.25	28.11	26.56	26.61	26.35	24.06
SwinIR [19]	ICCV 2021		32.98	29.12	26.76	29.59	26.77	25.00	28.62	26.56	25.18	27.05	23.86	22.30
DASR [36]	CVPR 2021	$\times 3$	34.08	33.57	31.15	29.99	28.66	28.42	28.90	28.62	28.13	27.36	26.86	25.95
HAT [5]	CVPR 2023		33.04	29.13	26.76	29.364	26.79	25.00	28.69	26.57	25.18	27.21	23.88	22.30
DSAT [22]	TMM 2024		34.56	33.77	31.96	**30.48**	**30.17**	28.98	29.22	29.10	28.24	**28.38**	27.88	**26.67**
DAFST (ours)			**34.57**	**34.01**	**32.27**	30.39	30.12	**29.06**	**29.25**	**29.18**	**28.30**	28.27	**27.89**	26.65
Method	Source	Scale	Set5			Set14			BSD100			Urban100		
			1.2	2.4	3.6	1.2	2.4	3.6	1.2	2.4	3.6	1.2	2.4	3.6
Bicubic			27.30	25.12	23.40	25.24	23.83	22.57	25.42	24.20	23.15	22.68	21.62	20.65
RCAN [52]	ECCV 2018		30.26	26.72	24.66	27.48	24.93	23.41	26.89	25.09	23.93	24.71	22.25	20.99
SRMD [50]	CVPR 2018		29.35	29.27	28.65	26.15	26.20	26.17	26.15	26.15	26.14	24.11	24.10	24.08
IKC [11]	CVPR 2019		31.77	30.56	29.23	28.45	28.16	26.81	27.43	27.27	26.33	25.63	25.00	24.06
SwinIR [19]	ICCV 2021		30.35	26.73	24.67	27.54	24.94	23.42	26.92	25.10	23.94	24.82	22.27	20.99
DASR [36]	CVPR 2021	$\times 4$	31.92	31.75	30.59	28.45	28.28	27.45	27.51	27.43	26.83	25.69	25.44	24.66
DAN [26]	Arxiv 2021		32.22	31.98	**30.94**	28.65	**28.54**	**27.69**	27.66	27.58	26.95	26.21	**25.97**	**25.08**
DSSR [17]	TMM 2022		32.26	32.09	30.89	26.68	**28.54**	**27.69**	27.65	27.58	26.95	26.09	25.83	24.97
HAT [5]	CVPR 2023		30.39	26.72	24.67	27.57	24.94	23.42	26.96	25.10	23.94	24.90	22.26	20.99
DSAT [22]	TMM 2024		**32.51**	32.00	30.31	28.67	28.50	27.51	**27.77**	**27.66**	26.98	**26.43**	25.95	24.89
DAFST (ours)			32.44	**32.25**	30.43	**28.78**	28.53	27.53	27.70	27.61	**27.02**	26.21	25.77	24.86

L1 loss. Like most image super-resolution works, we optimize our super-resolution network parameters by minimizing the L_1 pixel loss.

$$\mathcal{L}_{\text{SR}} = \left\| \mathbf{I}^{\text{HR}} - \mathbf{I}^{\text{SR}} \right\|_1 . \tag{16}$$

Total loss. The total loss function is defined as:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{SR}} + \mathcal{L}_{\text{degrade}}. \tag{17}$$

We optimize the network parameters of our degradation-aware frequency-separated Transformer by minimizing the total loss function. Our training strategy consists of two phases. We only train the degradation encoder for the first stage using $\mathcal{L}_{\text{degrade}}$ and then train the entire network using the $\mathcal{L}_{\text{Total}}$ in the second stage. This approach allows for focused optimization of degradation representations before integrating them into the whole model.

4. Experiments

4.1. Experimental setup

Dataset. Unless otherwise specified, all the compared methods were trained on a dataset combining the DIV2K [34] training set and the Flickr2K [21] dataset. We evaluated using benchmark super-resolution datasets, including Set5 [1], Set14 [46], BSD100 [27], and Urban100 [14].

Degradation settings. The degradation model of the LR image can be described as $\mathbf{I}^{\text{LR}} = (\mathbf{I}^{\text{HR}} \otimes \mathbf{k}_{(\lambda_1,\lambda_2,\Theta)}) \downarrow_r + \mathbf{n}_\epsilon$. According to the equation, LR images were synthesized for training and testing. The $\otimes$ denotes the convolution operation, the $\mathbf{k}_{(\lambda_1,\lambda_2,\Theta)}$ denotes the blur kernel controlled by $(\lambda_1, \lambda_2, \Theta)$; the noise $\mathbf{n}_\epsilon$ signifies additional random Gaussian noise with intensity ϵ; and $(\cdot) \downarrow_r$ indicates the downsampling operation with a scaling factor of r. The Gaussian kernel size is consistently set at 21×21. As the

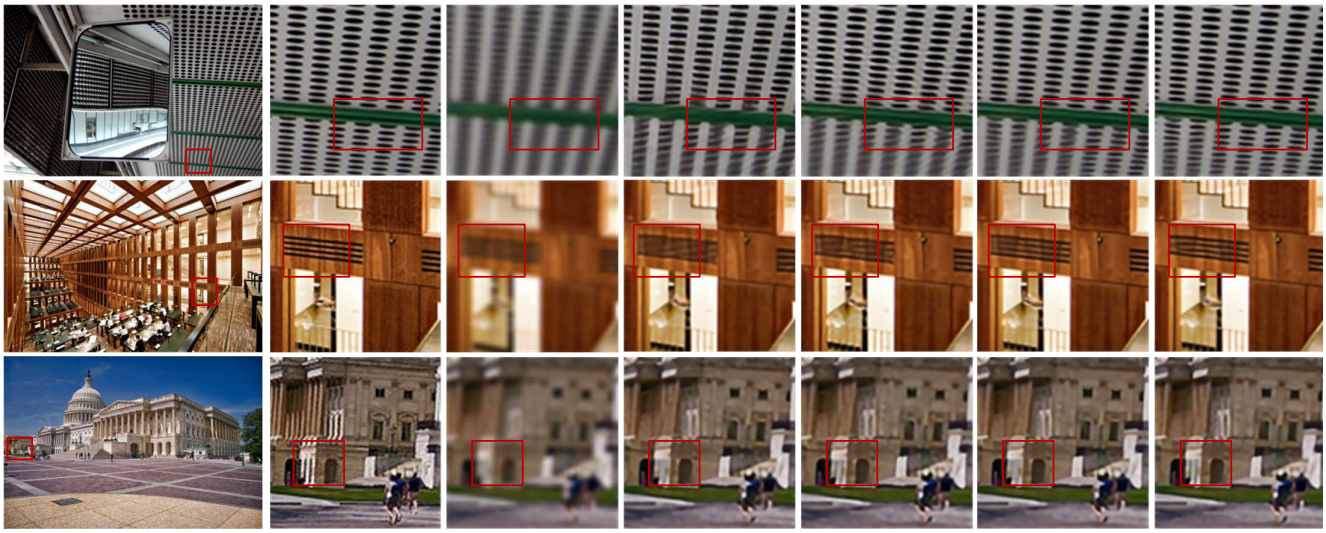


Figure 3: Visual results of compared methods and our DAFST on Urban100 dataset at a $\times 2$ SR setting using isotropic Gaussian kernel degradation with $\sigma = 1.2$. Our results show clearer details.

blur and noise intensities are unknown, the degree of degradation was adjusted by randomly sampling their hyperparameters, resulting in a diverse range of image degradations.

We conducted experiments under various degradation settings. In Subsection 4.2, we evaluated ours and the existing state-of-the-art (SOTA) methods under the condition of isotropic Gaussian kernel degradation with kernel widths σ (it is equivalent to $(\lambda_1 = \sigma, \lambda_2 = \sigma, \Theta = 0)$) within the range of [0.2, 2.0] for $\times 2$, [0.2, 3.0] for $\times 3$ and [0.2, 4.0] for $\times 4$ super-resolution training, respectively.

In Subsection 4.3 and Subsection 4.4, we further evaluated compared methods under more diverse conditions employing anisotropic Gaussian kernels and random noise to induce image degradation for $\times 4$ super-resolution. These anisotropic kernels are described by eigenvalues $\lambda_1, \lambda_2 \sim U(0.2, 4.0)$ and a randomly determined rotation angle $\Theta \sim U(0, \pi)$. Noise intensity ϵ is varied within the range of [0, 25]. If not specified otherwise, the rotation angle Θ and noise intensity ϵ of the blur kernel is 0 during testing.

Implementation. Our DAFST network consists of two depthwise convolution blocks (DWCB) and two frequency-separated self-attention groups (FSAG). The first group has six frequency-separated self-attention blocks (FSAB), and the second group has two. The hidden layer dimensionality is set to 180, with six attention heads per FSAB. Each FSAB is configured with five high-frequency ($O_1 = 5$) and one low-frequency ($O_2 = 1$) attention heads.

Additionally, for each image, we extracted one 48×48 sized query patch and three positive patches with different sizes (i.e., 32×32, 48×48, and 64×64, respectively) to improve the diversity of multi-patch.

During training, we used a batch size of 32 and applied data augmentation techniques, including 50% random vertical and horizontal flipping, as well as 50% random $90°$ rotations. We employed the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The training began with an initial learning rate of 10^{-4}, halved every 250 epochs. Our training has two phases: first, we trained the degradation encoder in isolation for 200 epochs with a learning rate of 10^{-3}, then trained the entire network using an additional 1000 epochs. All experiments were conducted on an NVIDIA GeForce RTX 4090 GPU.

4.2. Comparison on noise-free degradation

We compared our model under noise-free conditions using isotropic Gaussian kernels of different widths with bicubic interpolation, CNN-based methods (RCAN [52], SRMD [50], IKC [11], DASR [36], DAN [26], and DSSR [17]), and Transformer-based methods (SwinIR [19], HAT [5], and DSAT [22]), on Set5, Set14, BSD100, and Urban100 datasets.

Table 1 provides a quantitative comparison of our method against SOTA approaches across various kernel widths on four benchmark datasets. While all methods recover details well, performance notably drops for $\times 4$ SR compared to $\times 2$ SR, highlighting the greater challenge of $\times 4$ SR due to reduced contextual information. CNN-based methods, such as DSSR, have limited receptive fields, which hinders their ability to capture global features, resulting in a lack of competitiveness in $\times 2$ SR compared to $\times 4$ SR. Transformer-based methods, like DSAT, excel at extracting global features but may neglect finer details

Table 2: Quantitative comparison of $\times 4$ super-resolution on the Set14 dataset. The eigenvalues and rotation values (λ_1, λ_2, Θ) of anisotropic Gaussian kernels as well as the noise intensity (ϵ) are given for each column. **Bold** and underline represent the best and second-best performance, respectively.

Method	ϵ	 (2.0, 0.2, 0)	 (2.0, 1.0, 10)	 (3.5, 1.5, 30)	 (3.5, 2.0, 45)	 (3.5, 2.0, 90)	 (4.0, 1.5, 120)	 (4.0, 2.0, 135)	 (4.0, 3.0, 160)	 (4.0, 4.0, 180)
DnCNN + RCAN [52]	0	26.44	26.22	24.48	24.23	24.29	24.19	23.9	23.42	23.01
	5	26.10	25.90	24.29	24.07	24.14	24.02	23.74	23.31	22.92
	10	25.65	25.47	24.05	23.84	23.92	23.8	23.54	23.14	22.77
DnCNN + IKC [11]	0	27.71	27.78	27.11	27.02	26.93	26.65	26.5	26.01	25.33
	5	26.91	26.80	24.87	24.53	24.56	24.40	24.06	23.53	23.06
	10	26.16	26.09	24.55	24.33	24.35	24.17	23.92	23.43	23.01
DnCNN + DCLS [25]	0	27.56	27.49	26.32	25.99	25.88	26.03	25.70	24.65	23.95
	5	26.20	26.02	24.44	24.21	24.28	24.14	23.88	23.40	22.98
	10	25.47	25.33	24.06	23.87	23.91	23.79	23.58	23.16	22.78
DASR [36]	0	27.99	27.97	27.53	27.45	27.43	27.22	27.19	26.83	26.21
	5	27.25	27.18	26.37	26.16	26.09	25.96	25.85	25.52	25.04
	10	26.57	26.51	25.64	25.47	25.43	25.31	25.16	24.80	24.43
HAT-Real [5]	0	25.82	25.86	25.44	24.95	24.92	24.57	24.61	24.54	24.38
	5	25.74	25.80	25.21	24.81	24.78	24.49	24.46	24.20	23.93
	10	25.33	25.30	24.57	25.35	24.44	24.24	24.12	23.76	23.51
DSAT [22]	0	**28.34**	**28.34**	**27.78**	**27.68**	27.68	**27.37**	27.25	26.98	26.47
	5	27.55	27.47	26.59	26.43	26.43	26.31	26.14	25.80	25.36
	10	26.83	26.74	25.87	25.71	25.71	25.60	25.44	25.10	24.70
DAFST (ours)	0	28.29	28.31	27.76	27.66	**27.78**	**27.37**	**27.28**	**27.10**	**26.57**
	5	**27.68**	**27.65**	**26.68**	**26.51**	**26.52**	**26.40**	**26.29**	**25.92**	**25.46**
	10	**26.94**	**26.85**	**25.94**	**25.75**	**25.72**	**25.63**	**25.49**	**25.17**	**24.77**

critical for super-resolution. Methods like HAT, using hybrid attention mechanisms and larger attention windows, achieved significant performance in non-blind SR tasks. However, this enhancement did not improve the model's performance when faced with images of unknown degradation. In contrast, our low-frequency branch effectively captures global features with a larger receptive field, while the high-frequency branch enhances the generation of fine details. Our method demonstrates superior performance across all datasets for the SR tasks $\times 2$ and $\times 3$ and competitive results for $\times 4$ SR. The relatively modest performance in $\times 4$ SR is because the input image size is significantly small in $\times 4$ SR, which reduces the amount of useful information available for inference. This limitation is a common challenge shared by the compared methods as well.

Fig. 3 presents a visual comparison of the methods on the Urban100 dataset for the $\times 2$ SR task. The DSAT, guided by learned degradation representations, delivers impressive results. Constrained by multi-patch contrastive learning, our DAFST produces clearer textures and sharper edges than other methods. This multi-patch contrastive learning allows for extracting more discriminative degradation features, and the combination of low and high-frequency feature extraction captures both global and local visual details at larger kernel widths.

4.3. Comparison on general degradation

We conducted comparisons under general degradation conditions using anisotropic Gaussian kernels with added noise. We employed nine different anisotropic blur kernels and tested them under different noise intensities ϵ of 0, 5, and 10, respectively. Since RCAN [52], IKC [11], and DCLS [25] are not specifically designed to handle noise degradation, we incorporated a DnCNN [49] model as the pre-processing for fair comparisons.

Table 2 presents the quantitative results of the compared methods on the Set14 dataset. CNN-based approaches like RCAN and IKC demonstrate limited performance under complex degradation conditions. While IKC performs better, it relies on iterative estimation, making it time-consuming. HAT-Real is a version of HAT [5] trained according to ESRGAN's training settings [38] under multi-level degradation and noise conditions. Unlike our degradation-aware approach, it uses a generative adversarial network for training, which does not effectively distinguish between different levels of degradation. As a result, its performance on images with lower levels of degradation is limited. Implicit degradation representation-based methods, like DSAT, leverage contrastive learning to automatically learn degradation factors and demonstrate impressive performance. However, their reliance on a single view of the image during degradation learning may limit the learning effectiveness of degradation representations. In contrast, our DAFST employs multi-patch contrastive learning, allowing the model to learn more robust and discriminative degradation representations. This approach outperforms DCLS by over 1.0 dB across all degradation scenarios. Our method outperforms DSAT across most degradation scenarios. The multi-patch degradation learning provides clear guidance information to guide the subsequent super-resolution, and our high-frequency and

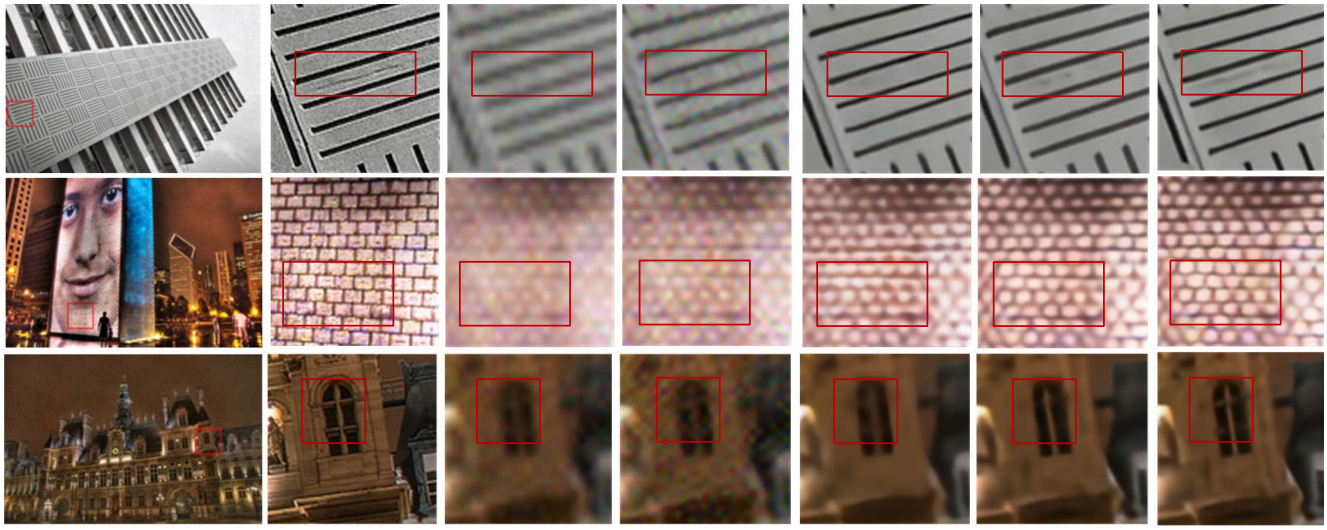


Figure 4: Visual results on the Urban100 dataset at a $\times 4$ SR setting using isotropic Gaussian kernel degradation with $\sigma = 1.8$ and $\epsilon = 5$. Our model outperforms others in preserving details in noisy images.

Table 3: Quantitative comparisons at $\times 4$ SR under varying noise intensities (ϵ) and kernel widths (σ).

Method	ϵ	Urban100 (PSNR/SSIM)			BSD100 (PSNR/SSIM)		
		$\sigma = 1$	$\sigma = 2$	$\sigma = 4$	$\sigma = 1$	$\sigma = 2$	$\sigma = 4$
DASR [36]	0	25.189/0.7527	24.777/0.7334	23.480/0.6667	27.315/0.7228	27.016/0.7094	26.060/0.6549
	5	24.823/0.7323	24.069/0.6952	22.456/0.6096	26.827/0.6932	26.209/0.6590	24.849/0.5920
	10	24.395/0.7123	23.595/0.6715	21.994/0.5857	26.269/0.6660	25.623/0.6313	24.335/0.5702
DSAT [22]	0	**25.684/0.7727**	**25.428/0.7606**	23.764/0.6720	**27.494/0.7318**	27.389/0.7244	26.229/0.6625
	5	25.364/0.7555	24.735/0.7267	22.931/0.6344	26.992/0.7031	26.458/0.6737	25.010/0.5991
	10	24.755/0.7326	24.070/0.6911	22.348/0.6063	26.380/0.6734	24.744/0.6413	24.505/0.5776
DAFST (ours)	0	25.663/0.7715	25.415/0.7593	**23.806/0.6809**	27.473/0.7300	**27.407/0.7242**	**26.288/0.6631**
	5	**25.398/0.7551**	**24.828/0.7264**	**23.023/0.6385**	**27.070/0.7066**	**26.591/0.6787**	**25.192/0.6094**
	10	**24.856/0.7302**	**24.121/0.6958**	**22.425/0.6080**	**26.452/0.6757**	**25.868/0.6439**	**24.611/0.5833**

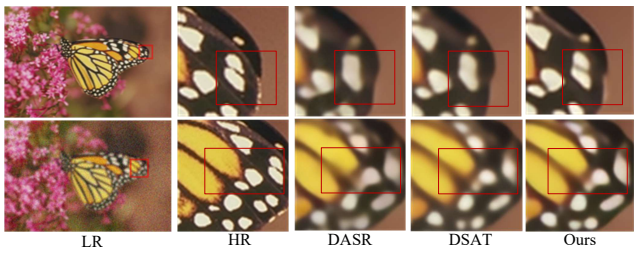


Figure 5: Visual comparison on the Set14 dataset. The settings of the anisotropic Gaussian blur kernel of the first row are $(\lambda_1 = 2.0, \lambda_2 = 0.2, \Theta = 0)$ and with $\epsilon = 0$; the settings of the blur kernel of the second are $(\lambda_1 = 4.0, \lambda_2 = 4.0, \Theta = 180)$ and with $\epsilon = 10$.

low-frequency branches and smaller attention windows help capture global and local details with high efficiency.

Fig. 4 illustrates the visual results of compared methods on the images from the Urban100 dataset with anisotropic Gaussian kernels and noise. The results reveal that DAN, which excelled under noise-free conditions, significantly deteriorates under noisy conditions. In contrast, our model adapts to more complex degradation scenarios, exhibiting clearer reconstruction results.

Fig. 5 presents qualitative comparisons with methods using implicit degradation representations, such as DASR and DSAT, on the Set14 dataset. Our method demonstrates advantages in the accuracy and clarity of detailed textures, further demonstrating its robust adaptability.

4.4. More comparisons

To further validate the effectiveness of our model, we compared it with methods using implicit degradation representations, like DASR and DSAT, on the Urban100 and BSD100 datasets. Since DSAT did not provide pre-trained

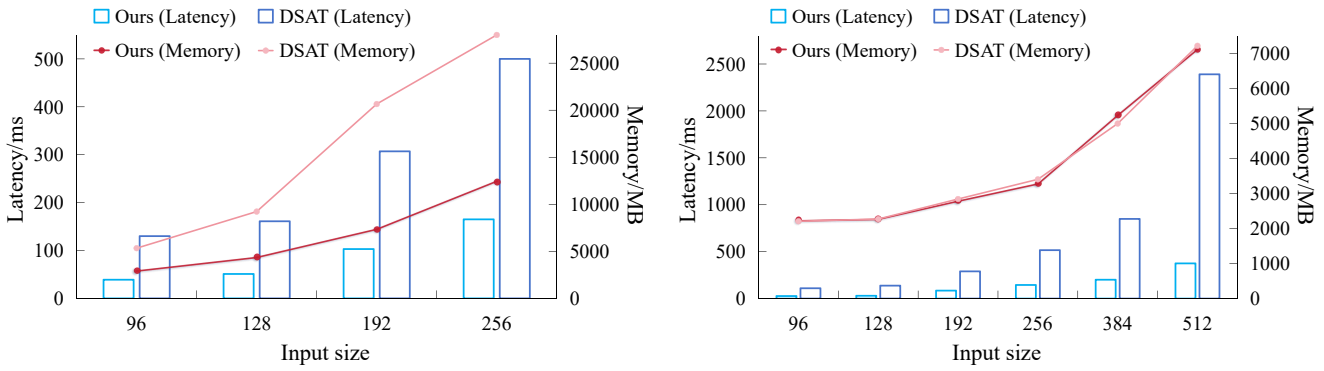


Figure 6: Inference speeds and memory usages of DSAT and our method under various image sizes in training (left) and testing (right) phases. When training at a resolution of 256×256, DSAT encountered an out-of-memory issue, while our method supports higher resolutions.

Table 4: Ablation study of proposed components on the Set14 dataset. The kernel widths of isotropic Gaussian kernels (σ) and noise intensities (ϵ) are given for each column. The ratio means the numbers of high-to-low frequency attention heads. The DR represents the degradation representations learning, and the DWMLP means the DWMLP block.

Method	Ratio	DR	DWMLP	(1.2, 0) PSNR/SSIM	(1.2, 5) PSNR/SSIM	(1.2, 10) PSNR/SSIM	(2.4, 0) PSNR/SSIM	(2.4, 5) PSNR/SSIM	(2.4, 10) PSNR/SSIM	(3.6, 0) PSNR/SSIM	(3.6, 5) PSNR/SSIM	(3.6, 10) PSNR/SSIM
Model-I	6:0	✓	✓	28.332/0.7714	27.844/0.7490	27.129/0.7205	28.043/**0.7574**	26.960/0.7101	26.122/0.6777	26.880/0.7049	25.869/0.6674	25.087/0.6386
Model-II	3:3	✓	✓	28.315/0.7714	27.834/0.7487	27.097/0.7193	27.978/0.7540	26.943/0.7101	26.095/0.6770	26.843/0.7015	25.890/0.6668	25.096/0.6383
Model-III	1:5	✓	✓	28.304/0.7694	27.829/0.7488	27.084/0.7199	27.952/0.7513	26.966/0.7108	26.109/0.6767	26.860/0.7036	25.880/0.6670	25.100/0.6382
Model-IV	5:1	×	✓	28.265/0.7704	27.766/0.7477	27.043/0.7190	27.889/0.7437	26.869/0.7068	26.012/0.6755	26.764/0.6959	25.810/0.6635	25.041/0.6362
Model-V	5:1	✓	×	28.344/0.7702	27.838/0.7479	27.092/0.7196	28.041/0.7555	26.947/0.7099	26.115/0.6774	26.972/**0.7070**	25.891/0.6658	25.130/0.6384
Model-VI	5:1	×	×	28.221/0.7701	27.738/0.7455	27.029/0.7170	27.836/0.7432	26.861/0.7077	26.002/0.6727	26.809/0.6978	25.618/0.6549	24.990/0.6326
Model-VII (ours)	5:1	✓	✓	**28.423/0.7728**	**27.904/0.7505**	**27.144/0.7215**	**28.101**/0.7531	**27.035/0.7118**	**26.168/0.6795**	**27.022**/0.7040	**25.916/0.6684**	**25.155/0.6405**

weights, we retrained it according to the method described by the authors. Quantitative results shown in Table 3 indicate that our method also performs better on the Urban100 and BSD100 datasets.

Moreover, we show the memory consumption and inference speed, compared with the SOTA Transformer-based approach, DSAT. We specified the image size and input images to the model for $\times 4$ super-resolution. The time is averaged over 1000 samples. As shown in Fig. 6, our model consumes significantly less memory and higher speed during training than DSAT. Our model can be trained on datasets with large sizes to tap into the model's potential further. In detail, the network of DSAT has 15.64 million (M) of parameters, while our proposed method has a significantly reduced parameter number of 10.15M. Our degradation-aware frequency-separated attention mechanism utilizes smaller attention windows for high-frequency branches and compact visual embeddings for low-frequency branches to capture finer details and global structures for efficient BSR.

4.5. Ablation study

We conducted ablation studies to evaluate the effectiveness of our proposed components. Specifically, the number ratio of high-to-low frequency attention heads in Model-I is 6:0 ($O = 6$ and $\hat{O} = 0$), which is equivalent to standard window-based attention; Model-II and Model-III have ratios of 3:3 ($O = 3$ and $\hat{O} = 3$) and 1:5 ($O = 1$ and $\hat{O} = 5$), respectively; Model-IV removes the degradation representations learning (DR); Model-V removes the DWConv within DWMLP and Model-VI remove both DR and DWConv; Model-VII (our full model) utilizes a high-to-low frequency ratio of 5:1 ($O = 5$ and $\hat{O} = 1$) and incorporates both DR and DWMLP.

As shown in Table 4, the results indicate that Model-VII outperformed all other models across all evaluation metrics. Model-IV, which lacks implicit representation guidance, significantly decreases SR performance. Model-V, which does not utilize depthwise convolution to aggregate high and low-frequency attention branches, exhibits a slight decrease in quantitative scores. Experimental results show that the best results are achieved when the ratio of attention heads of high-to-low frequency branches is 5:1, indicating that more high-frequency information may be crucial in blind super-resolution tasks.

As shown in Table 5, we validated the effectiveness of the DWCB for shallow feature extraction. For a fair comparison, we created three variants with nearly identical parameters: one replaces the DWCB with two frequency-separated self-attention blocks (Model-VIII), an-

Table 5: Ablation study of different position encoding learning schemes on the Set14 dataset. The kernel widths (σ) of isotropic Gaussian kernels and noise levels (ϵ) are given for each column. The latency values are averaged over 1000 samples, each with an input resolution of 512×512.

Method	(1.2, 0) PSNR/SSIM	(1.2, 5) PSNR/SSIM	(1.2, 10) PSNR/SSIM	(2.4, 0) PSNR/SSIM	(2.4, 5) PSNR/SSIM	(2.4, 10) PSNR/SSIM	(3.6, 0) PSNR/SSIM	(3.6, 5) PSNR/SSIM	(3.6, 10) PSNR/SSIM	Latency
Model-VIII	28.314/0.7702	27.824/0.7485	27.097/0.7200	28.003/0.7540	26.953/0.7097	26.115/0.6775	26.975/**0.7084**	25.884/0.6660	25.130/0.6386	442ms
Model-IX	28.370/0.7717	27.822/0.7480	27.126/0.7200	28.014/0.7530	26.983/0.7093	**26.175**/0.6785	26.976/0.7082	**25.923**/0.6665	25.153/0.6393	467ms
Model-VII (ours)	**28.423 0.7728**	**27.904/0.7505**	**27.144/0.7215**	**28.101/0.7531**	**27.035/0.7117**	26.168/**0.6795**	**27.022**/0.7040	25.916/**0.6684**	**25.155/0.6405**	**375ms**

Table 6: Ablation study for different positive sample numbers on Set5 and Urban100 datasets (using PSNR). The scaling factor of super-resolution is $\times 2$; the kernel widths (σ) of isotropic Gaussian kernels are given for each column.

Positive samples	Set5				Urban100			
	0.0	0.5	1.0	1.5	0.0	0.5	1.0	1.5
1	38.082	38.111	37.946	36.897	32.225	**31.820**	**31.501**	30.367
3 (ours)	**38.130**	**38.141**	**37.968**	**36.931**	**32.325**	31.805	31.494	**30.417**
5	38.103	38.091	37.840	36.900	32.242	31.666	31.408	30.314
7	38.099	38.058	37.888	36.909	32.202	31.695	31.330	30.225

other replaces the DWCB with two frequency-separated self-attention blocks incorporating convolutional position encoding Model-IX, and Model-VII (our full model). Our method with the DWCB demonstrates the best quantitative performance while maintaining faster inference speed.

4.6. Analysis

Analysis on the number of positive samples. In our blind super-resolution task, contrastive learning efficacy exhibits a non-linear trend as the number of positive samples varies. As shown in Table 6, the optimal PSNR performance is achieved with three positive samples, followed by using one positive sample, while performance declines with five and seven samples. This phenomenon likely stems from the unique demands of capturing image degradation features in blind super-resolution. Three positive samples appear to achieve a good balance, maintaining sufficient feature variability while avoiding overfitting specific visual features. Thus, it enables the model to learn rich degradation representations while preserving sensitivity to specific patterns. Conversely, with five or seven samples, the model may overfit on over-averaged representations from the whole dataset, overlooking subtle yet crucial differences that play a pivotal role in high-quality image reconstruction.

Analysis on the window size of high-frequency attention branch. We further conducted an ablation on different sizes of local attention windows based on Model-VII. The results in Table 7 indicate that appropriately sized local attention windows are more beneficial for feature extraction modeling. Excessively small or large windows do not necessarily facilitate extracting high-frequency local features.

As shown in Fig. 7, we visualized the qualitative performance of the models on Set14. The results demonstrate that our method (Model-VII) exhibits high-fidelity visual results, with the reconstructed high-resolution images containing rich textural details.

Table 7: Ablation study of different window sizes for high-frequency attention branch ($\times 4$ BSR). The kernel width of isotropic Gaussian is $\sigma = 1.2$.

Window size	Set5 PSNR/SSIM	Set14 PSNR/SSIM	BSD100 PSNR/SSIM	Urban100 PSNR/SSIM
2×2	31.864/**0.8907**	28.403/0.7711	27.455/**0.7294**	25.620/0.7690
4×4 (ours)	**31.948**/ 0.8903	**28.423/0.7728**	**27.460**/0.7272	**25.649**/0.7692
8×8	31.857/0.8893	28.377/0.7707	27.443/0.7264	25.627/**0.7696**

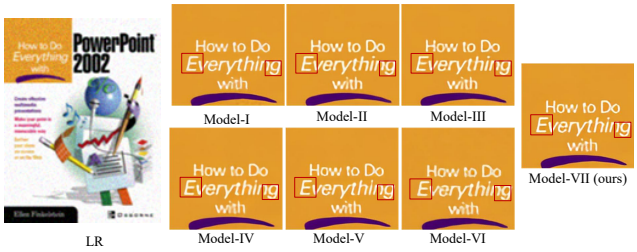
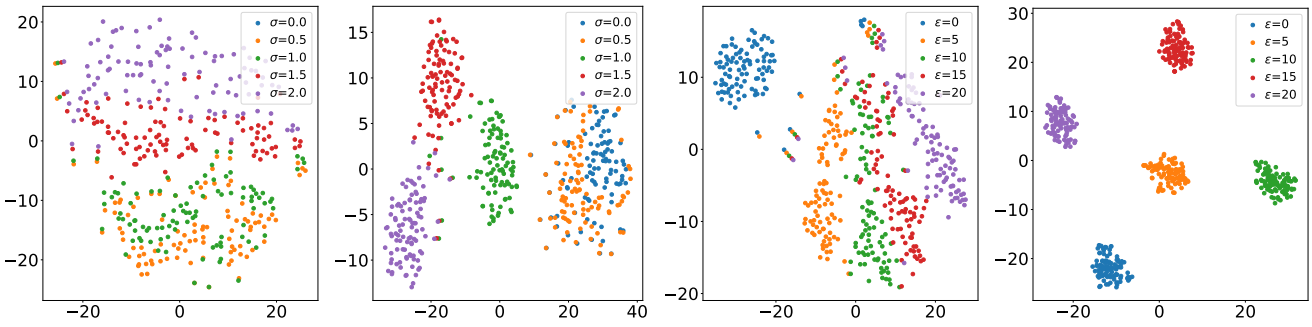


Figure 7: Visualization of $\times 4$ super-resolution ablation study under the conditions of $\sigma = 2.4$ and $\epsilon = 5$. The SR image generated by Model-VII is visually more appealing, with clearer edge details in letters.

Analysis on degradation representation learning. Here, we further analyze the effectiveness of the degradation representation learning by visualizing the projected degradation representations using degradation encoders in Model-IV and Model-VII (our full model), respectively. Note that the Model-IV was trained without the multi-patch contrastive learning. We used the BSD100 dataset to create LR images with different degradation degrees and input them into Model-IV and Model-VII to obtain embeddings from the degradation encoder. We then visualized these representations using the T-SNE [35].

Fig. 8 shows that embeddings using our Model-VII with degradation learning can be distinctively clustered when faced with different kernel widths or varying levels of noise degradation. The performance of Model-IV and Model-VII on images degraded under different kernel widths and noise



(a) Model-IV (b) Model-VII (ours) (c) Model-IV (d) Model-IV (ours)

Figure 8: Visualization of degradation representations under different kernel widths (σ) and noise levels (ϵ) using degradation encoder. (a) and (c) are without degradation learning. (b) and (d) are with our degradation learning.

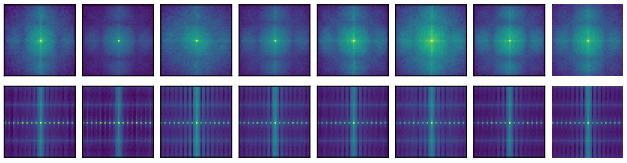


Figure 9: The frequency amplitude (48×48) of the first eight output channels of the high branch (upper) and low branch (lower), in the last FSAB. The amplitude is averaged over 32 samples. The lighter the color, the higher the amplitude. Pixels closer to the center represent lower frequencies, and vice versa.

intensities. Table 4 also proves that the learning of degradation representations indeed helps our encoder learn discriminative representations to provide useful guidance for better blind super-resolution.

Visualization of frequency branches. In Fig. 9, we visualize the amplitude of frequency components by applying the Fast Fourier Transform (FFT) [29] to feature maps of the high and low-frequency self-attention branches separately. Some periodic spectral changes might be caused by the applied degrading blur kernel. Our visualization results indicate that the high-frequency branch captures more high-frequency information. In Fig. 3 first row, our model correctly restores the shape of the holes. In contrast, the low-frequency branch primarily focuses on low-frequency information. In Fig. 4 first and third rows, our method successfully recovers the structure of buildings. These results demonstrate the effectiveness of our proposed degradation-aware frequency-separated Transformer for high-quality blind super-resolution.

5. Conclusion

In this paper, we have presented a novel degradation-aware frequency-separated Transformer for blind super-resolution. Our approach effectively captures discriminative degradation representations through multi-patch contrastive learning. The proposed frequency-separated Transformer leverages degradation representations to efficiently extract both local details and global structural features via high- and low-frequency branches, combining fine-grained details with global context. Extensive comparisons, ablation studies, and analyses demonstrated the superior performance of our method and the effectiveness of the proposed components for blind super-resolution.

Our method has some limitations. First, like most Transformer-based methods, it faces significant memory and computational overhead when performing ultra-high-resolution images. Second, for extremely low-resolution images, performance is affected due to limited available information. Expanding the range of frequency representations would enhance the model's learning capacity. In the future, we aim to address these challenges by incorporating a broader range of frequencies. We would like to explore the application of our degradation-aware frequency-separated Transformer in other computer vision tasks, such as image recognition and detection.

References

[1] M. Bevilacqua, A. Roumy, C. Guillemot, and M. Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *BMVC*, pages 1–10. BMVA, 2012. 7

[2] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *NeurIPS*, 33:9912–9924, 2020. 2

[3] A. Chadha, J. Britto, and M. M. Roja. iseebetter: Spatio-temporal video super-resolution using recurrent generative back-projection networks. *CVMJ*, 6:307–317, 2020. 1

[4] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao. Pre-trained image processing transformer. In *CVPR*, pages 12299–12310. IEEE, 2021. 1

[5] X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong. Activating more pixels in image super-resolution transformer. In *CVPR*, pages 22367–22377. IEEE, 2023. 7, 8, 9

[6] T. Dai, J. Cai, Y. Zhang, S. Xia, and L. Zhang. Second-order attention network for single image super-resolution. In *CVPR*, pages 11065–11074. IEEE, 2019. 1, 2

[7] C. Dong, C. C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, volume 8692, pages 184–199. Springer, 2014. 2

[8] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE TPAMI*, 38(2):295–307, 2015. 2

[9] C. Dong, C. C. Loy, and X. Tang. Accelerating the super-resolution convolutional neural network. In *ECCV*, volume 9906, pages 391–407. Springer, 2016. 2, 4

[10] M. Fritsche, S. Gu, and R. Timofte. Frequency separation for real-world super-resolution. In *ICCVW*, pages 3599–3608. IEEE, 2019. 2

[11] J. Gu, H. Lu, W. Zuo, and C. Dong. Blind super-resolution with iterative kernel correction. In *CVPR*, pages 1604–1613. IEEE, 2019. 3, 7, 8, 9

[12] H. Guo, J. Li, T. Dai, Z. Ouyang, X. Ren, and S.-T. Xia. Mambair: A simple baseline for image restoration with state-space model. In *ECCV*, volume 15076, pages 222–241. Springer, 2024. 2

[13] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9726–9735. IEEE, 2020. 6

[14] J. Huang, A. Singh, and N. Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, pages 5197–5206. IEEE, 2015. 7

[15] J. Kim, J. K. Lee, and K. M. Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, pages 1646–1654. IEEE, 2016. 2

[16] P. H. Le-Khac, G. Healy, and A. F. Smeaton. Contrastive representation learning: A framework and review. *IEEE ACCESS*, 8:193907–193934, 2020. 6

[17] F. Li, Y. Wu, H. Bai, W. Lin, R. Cong, and Y. Zhao. Learning detail-structure alternative optimization for blind super-resolution. *IEEE TMM*, 25:2825–2838, 2023. 1, 3, 7, 8

[18] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu. Feedback network for image super-resolution. In *CVPR*, pages 3867–3876. IEEE, 2019. 2

[19] J. Liang, J. Cao, G. Sun, K. Zhang, L. V. Gool, and R. Timofte. Swinir: Image restoration using swin transformer. In *ICCVW*, pages 1833–1844. IEEE, 2021. 1, 2, 6, 7, 8

[20] J. Liang, H. Zeng, and L. Zhang. Efficient and degradation-adaptive network for real-world image super-resolution. In *ECCV*, volume 13678, pages 574–591. Springer, 2022. 2

[21] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee. Enhanced deep residual networks for single image super-resolution. In *CVPR*, pages 1132–1140. IEEE, 2017. 7

[22] Q. Liu, P. Gao, K. Han, N. Liu, and W. Xiang. Degradation-aware self-attention based transformer for blind image super-resolution. *IEEE TMM*, 26:7516–7528, 2024. 1, 3, 7, 8, 9, 10

[23] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *CVPR*, pages 10012–10022. IEEE, 2021. 3

[24] W. Lu, J. Wang, T. Wang, K. Zhang, X. Jiang, and H. Zhao. Visual style prompt learning using diffusion models for blind face restoration. *PR*, 161:111312, 2025. 1

[25] Z. Luo, H. Huang, L. Yu, Y. Li, H. Fan, and S. Liu. Deep constrained least squares for blind image super-resolution. In *CVPR*, pages 17621–17631. IEEE, 2022. 3, 9

[26] Z. Luo, Y. Huang, S. Li, L. Wang, and T. Tan. End-to-end alternating optimization for blind super resolution. *arXiv preprint arXiv:2105.06878*, 2021. 3, 7, 8

[27] D. R. Martin, C. 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 *ICCV*, pages 416–425. IEEE, 2001. 7

[28] Z. Pan, J. Cai, and B. Zhuang. Fast vision transformers with hilo attention. *NeurIPS*, 35:14541–14554, 2022. 1, 2

[29] Y. Rao, W. Zhao, Z. Zhu, J. Lu, and J. Zhou. Global filter networks for image classification. *NeurIPS*, 34:980–993, 2021. 13

[30] M. She, W. Mao, H. Shi, and Z. Wang. S 2 r: Exploring a double-win transformer-based framework for ideal and blind super-resolution. In *ICANN*, pages 522–537. Springer, 2023. 1

[31] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, pages 1874–1883. IEEE, 2016. 2

[32] A. Shocher, N. Cohen, and M. Irani. "zero-shot" super-resolution using deep internal learning. In *CVPR*, pages 3118–3126. IEEE, 2018. 3

[33] J. W. Soh, S. Cho, and N. I. Cho. Meta-transfer learning for zero-shot super-resolution. In *CVPR*, pages 3513–3522. IEEE, 2020. 3

[34] 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 *CVPR*, pages 1110–1121. IEEE, 2017. 7

[35] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008. 12

[36] L. Wang, Y. Wang, X. Dong, Q. Xu, J. Yang, W. An, and Y. Guo. Unsupervised degradation representation learning for blind super-resolution. In *CVPR*, pages 10581–10590. IEEE, 2021. 1, 3, 7, 8, 9, 10

[37] T. Wang, G. Tao, W. Lu, K. Zhang, W. Luo, X. Zhang, and T. Lu. Restoring vision in hazy weather with hierarchical contrastive learning. *PR*, 145:109956, 2024. 1

[38] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy. ESRGAN: enhanced super-resolution generative adversarial networks. In *ECCV*, volume 11133, pages 63–79. Springer, 2018. 9

[39] Y. Wang, Y. Liu, S. Zhao, J. Li, and L. Zhang. Camixersr: Only details need more "attention". *arXiv preprint arXiv:2402.19289*, 2024. 2

[40] Y. Wei, S. Gu, Y. Li, R. Timofte, L. Jin, and H. Song. Unsupervised real-world image super resolution via domain-distance aware training. In *CVPR*, pages 13385–13394. IEEE, 2021. 3

[41] B. Xia, Y. Zhang, Y. Wang, Y. Tian, W. Yang, R. Timofte, and L. V. Gool. Knowledge distillation based degradation estimation for blind super-resolution. In *ICLR*. OpenReview.net, 2023. 3

[42] J. Xu, M. Yuan, D.-M. Yan, and T. Wu. Deep unfolding multi-scale regularizer network for image denoising. *CVMJ*, 9:335–350, 2023. 1

[43] Y. Xu, S. R. Tseng, Y. Tseng, H. Kuo, and Y. Tsai. Unified dynamic convolutional network for super-resolution with variational degradations. In *CVPR*, pages 12493–12502. IEEE, 2020. 2

[44] J. Yang, C. Li, P. Zhang, X. Dai, B. Xiao, L. Yuan, and J. Gao. Focal self-attention for local-global interactions in vision transformers. *arXiv preprint arXiv:2107.00641*, 2021. 6

[45] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue, and Q. Liao. Deep learning for single image super-resolution: A brief review. *IEEE TMM*, 21(12):3106–3121, 2019. 2

[46] R. Zeyde, M. Elad, and M. Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces*, pages 711–730. Springer, 2012. 7

[47] K. Zhang, L. V. Gool, and R. Timofte. Deep unfolding network for image super-resolution. In *CVPR*, pages 3214–3223. IEEE, 2020. 2

[48] K. Zhang, Y. Li, W. Zuo, L. Zhang, L. Van Gool, and R. Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE TPAMI*, 44(10):6360–6376, 2021. 2

[49] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE TIP*, 26(7):3142–3155, 2017. 9

[50] K. Zhang, W. Zuo, and L. Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *CVPR*, pages 3262–3271. IEEE, 2018. 2, 7, 8

[51] K. Zhang, W. Zuo, and L. Zhang. Deep plug-and-play super-resolution for arbitrary blur kernels. In *CVPR*, pages 1671–1681. IEEE, 2019. 2

[52] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, volume 11211, pages 294–310. Springer, 2018. 1, 2, 7, 8, 9

[53] Y. Zhang, Z. Tan, J. Yang, W. Huang, and Y. Yuan. Matrix information theory for self-supervised learning. *arXiv preprint arXiv:2305.17326*, 2023. 2