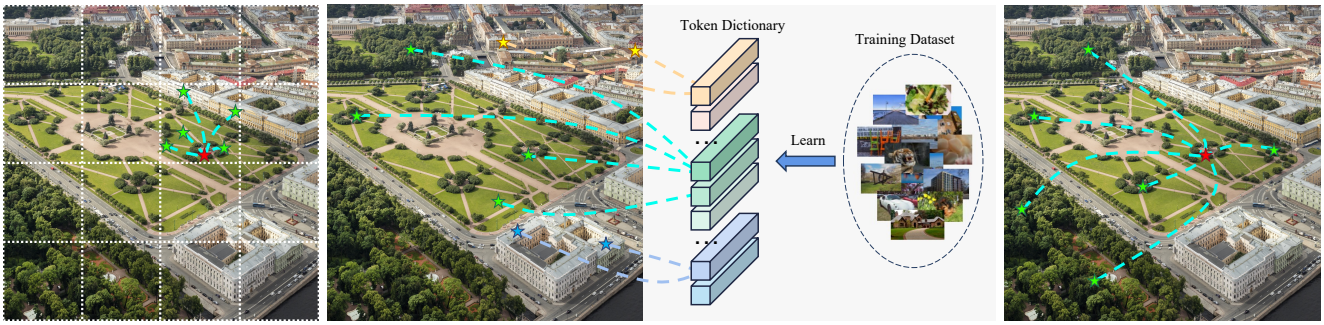
Transcending the Limit of Local Window: Advanced Super-Resolution Transformer with Adaptive Token Dictionary

Leheng Zhang[1] Yawei Li[2,3] Xingyu Zhou[1] Xiaorui Zhao[1] Shuhang Gu[1*]
[1]University of Electronic Science and Technology of China [2]Computer Vision Lab, ETH Zürich
[3]Integrated Systems Laboratory, ETH Zürich
{lehengzhang12, shuhanggu}@gmail.com

https://github.com/LabShuHangGU/Adaptive-Token-Dictionary



(a) window-based self-attention (b) token dictionary cross-attention (c) category-based self-attention

Figure 1. Three different kinds of attention mechanism: (a) window-based self-attention exploits tokens in the same local window to enhance image tokens; (b) our proposed token dictionary cross-attention leverages the auxiliary dictionary to summarize and incorporate global information to the image tokens; (c) our proposed category-based self-attention adopts category labels to divide image tokens.

Abstract

*Single Image Super-Resolution is a classic computer vision problem that involves estimating high-resolution (HR) images from low-resolution (LR) ones. Although deep neural networks (DNNs), especially Transformers for super-resolution, have seen significant advancements in recent years, challenges still remain, particularly in limited receptive field caused by window-based self-attention. To address these issues, we introduce a group of auxiliary **A**daptive **T**oken **D**ictionary to SR Transformer and establish an **ATD-SR** method. The introduced token dictionary could learn prior information from training data and adapt the learned prior to specific testing image through an adaptive refinement step. The refinement strategy could not only provide global information to all input tokens but also group image tokens into categories. Based on category partitions, we further propose a category-based self-attention mechanism designed to leverage distant but similar tokens for enhancing input features. The experimental results show that our method achieves the best performance on various single image super-resolution benchmarks.*

*corresponding author

1. Introduction

The task of single image super-resolution (SR) aims to recover clean high-quality (HR) images from a solitary degraded low-quality (LR) image. Since each LR image may correspond to a mass of possible HR counterparts, image SR is a classical ill-posed and challenging problem in the fields of computer vision and image processing. This practice is significant as it transcends the resolution and accuracy limitations of cost-effective sensors and improves images produced by outdated equipment.

The evolution of image super-resolution techniques has shifted from earlier methods like Markov random fields [14] and Dictionary Learning [39] to advanced deep learning approaches. The rise of deep neural networks, particularly convolutional neural networks (CNNs), marked a significant improvement in this field, with models effectively learning mapping functions from LR to HR images [9, 10, 17, 23, 43]. More recently, Transformer-based neural networks have outperformed CNNs in image super-resolution by employing self-attention mechanisms to better model long-range image structures [4, 21, 22].

Despite recent advances in image SR, several challenges remain unresolved. One major issue faced by SR transform-

2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-5300-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/CVPR52733.2024.00276

2575-7075/24/$31.00 ©2024 IEEE
DOI 10.1109/CVPR52733.2024.00276

Authorized licensed use limited to: ETH BIBLIOTHEK ZURICH. Downloaded on September 23,2025 at 13:50:29 UTC from IEEE Xplore. Restrictions apply.

ers is the balancing act between achieving satisfactory SR accuracy and managing increased computational complexity. Due to the quadratic computational complexity of the self-attention mechanism, previous methods [22, 24] have been forced to confine attention computation to local windows to manage computational load. However, this window-based method imposes a constraint on the receptive field, affecting the performance. Although recent studies [4, 21] indicate that expanding window size improves the receptive field and enhances SR performance, it exacerbates the curse of dimensionality. This issue underscores the need for an efficient method to effectively model long-range dependencies, without being constrained within local windows. Furthermore, conventional image SR often employs general-purpose computations that do not take the content of the image into account. Rather than employing a partitioning strategy based on rectangular local windows, opting for division according to the specific content categories of an image could be more beneficial to the SR process.

In our paper, we draw inspiration from classical dictionary learning in super-resolution to introduce a token dictionary, enhancing both cross- and self-attention calculations in image processing. This token dictionary offers three distinct benefits. **Firstly**, it enables the use of cross-attention to integrate external priors into image analysis. This is achieved by learning auxiliary tokens that encapsulate common image structures, facilitating efficient processing with linear complexity in proportion to the image size. **Secondly**, it enables the use of global information to establish long-range connection. This is achieved by refining the dictionary with activated tokens to summarize image-specific information globally through a reversed form of attention. **Lastly**, it enables the use of all similar parts of the image to enhance image tokens without being limited by local window partitions. This is achieved by content-dependent structural partitioning according to the similarities between image and dictionary tokens for category-based self-attention. These innovations enable our method to significantly outperform existing state-of-the-art techniques without substantially increasing the model complexity.

Our contributions can be summarized as follows:

- We introduce the idea of token dictionary, which utilizes a group of auxiliary tokens to provide prior information to each image token and summarize prior information from the whole image, effectively and efficiently in a cross-attention manner.
- We exploit our token dictionary to group image tokens into categories and break through boundaries of local windows to exploit long-range prior in a category-based self-attention manner.
- By combining the proposed token dictionary cross-attention and category-based self-attention, our model could leverage long-range dependencies effectively and achieve superior super-resolution results over the existing state-of-the-art methods.

2. Related Works

The past decade has witnessed numerous endeavors aimed at improving the performance of deep learning methods across diverse fields, including the single image super-resolution. Pioneered by SRCNN [10], which introduces deep learning to super-resolution with a straightforward 3-layer convolutional neural network (CNN), numerous studies have since explored various architectural enhancements to boost performance [9, 13, 17, 18, 23, 31–33, 43, 44]. VDSR [17] implements a deeper network, and DRCN [18] proposes a recursive structure. EDSR [23] and RDN [44] develop new residual blocks, further improving CNN capability in SR. Drawing inspiration from Transformer [35], Wang et al. [38] first integrates non-local attention block into CNN, validating the effects of attention mechanism in vision tasks. Following that, numerous advances in attention have emerged. CSNLN [31] makes use of non-local cross-scale attention to explore cross-scale feature correlations and mine self-exemplars in natural images. NLSA [32] further improves efficiency through sparse attention, which reduces the calculation between unrelated or noisy contents.

Recently, with the introduction of ViT [11] and its variants [7, 24, 37], the efficacy of pure Transformer-based models in image classification has been established. Based on this, IPT [3] makes a successful attempt to exploit the Transformer-based network for various image restoration tasks. Since then, a variety of techniques have been developed to enhance the performance of super-resolution transformers. This includes the implementation of shifted window self-attention by SwinIR [22] and CAT [5], group-wise multi-scale self-attention by ELAN [42], sparse attention by ART [41] and OmniSR [36], anchored self-attention by GRL [21], and more, all aimed at expanding the scope of receptive field to achieve better results. Furthermore, strategies such as pretraining on extensive datasets [20], employing ConvFFN [36], and utilizing large window sizes [4] have been employed to boost performance, indicating the growing adaptability and impact of Transformer-based approaches in the field of image SR.

In this paper, building upon the effectiveness of the attention mechanism in image SR, we propose two types of attention: token dictionary cross-attention (TDCA) to leverage external prior and adaptive category-based multi-head self-attention (AC-MSA) to model long-range dependencies. When synergized with window-based attention, our approach seamlessly integrates local, global, and external information, yielding promising outcomes in image super-resolution tasks.

Authorized licensed use limited to: ETH BIBLIOTHEK ZURICH. Downloaded on September 23,2025 at 13:50:29 UTC from IEEE Xplore. Restrictions apply.

3. Methodology

3.1. Motivation

In this subsection, we introduce the motivation of our approach. We first discuss how dictionary-learning-based SR methods utilize learned dictionary to provide supplementary information for image SR. Then, we analyze the attention operation and discuss its similarity to the coefficient calculation and signal reconstruction processes in dictionary-learning-based methods. Lastly, we discuss how these two methods motivate us to introduce an auxiliary token dictionary for enhancing both cross- and self-attention calculations in image processing.

Dictionary Learning for Image Super-Resolution. Before the era of deep learning, dictionary learning plays an important role in providing prior information for image SR. Due to the limited computational resources, conventional dictionary-learning-based methods divide image into patches for modeling image local prior. Denote $\boldsymbol{x} \in \mathbb{R}^d$ as a vectorized image patch in the low-resolution (LR) image. To estimate the corresponding high-resolution (HR) patch $\boldsymbol{y} \in \mathbb{R}^d$, Yang et al. [39] decompose the signal by solving the sparse representation problem:

$$\boldsymbol{\alpha}^* = argmin_{\boldsymbol{\alpha}} \|\boldsymbol{x} - \boldsymbol{D}_L \boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1 \qquad (1)$$

and reconstruct the HR patch with $\boldsymbol{D}_H \boldsymbol{\alpha}^*$; where $\boldsymbol{D}_L \in \mathbb{R}^{d \times M}$ and $\boldsymbol{D}_H \in \mathbb{R}^{d \times M}$ are the learned LR and HR dictionaries, and M is the number of atoms in the dictionary. Most of dictionary-learning-based SR methods [12, 39, 40] learn coupled dictionaries $\boldsymbol{D}_L$ and $\boldsymbol{D}_H$ to summarize the prior information from the external training dataset; several attempts [27, 28] have also been made to refine dictionary according to the testing image for better SR results.

Vision Transformer for Image Super-Resolution. Recently, Transformer-based approaches have pushed the state-of-the-art of many vision tasks to a new level. At the core of Transformer is the self-attention operation, which exploits similarity between tokens as weight to mutually enhance image features:

$$\text{Atten}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{SoftMax}\left(\boldsymbol{Q}\boldsymbol{K}^T / \sqrt{d}\right) \boldsymbol{V}; \qquad (2)$$

$\boldsymbol{Q} \in \mathbb{R}^{N \times d}$, $\boldsymbol{K} \in \mathbb{R}^{N \times d}$ and $\boldsymbol{V} \in \mathbb{R}^{N \times d}$ are linearly transformed from the input feature $\boldsymbol{X} \in \mathbb{R}^{N \times d}$ itself, N is the token number and d is the feature dimension. Due to the self-attentive processing philosophy, the large window size plays a critical role in modeling the internal prior of more patches. However, the complexity of self-attention computation increases quadratically with the number of input tokens, and different strategies including shift-window [8, 22, 24, 25], anchor attention [21], and shifted crossed attention [20] have been proposed to alleviate the limited window size issue of the Vision Transformer.

Advanced Cross&Self-Attention with Token Dictionary. After reviewing the above content, we found that the decomposition and reconstruction idea of dictionary learning-based image SR is similar to the process of self-attention computation. Specifically, the above method in Eq. (1) solves the sparse representation model to find similar LR dictionary atoms and reconstruct HR signal with the corresponding HR dictionary atoms; while attention-based methods use normalized inner product operation to determine attention weights to combine value tokens.

The above observation implies that the idea of dictionary learning can be easily incorporated into the Transformer framework for breaking the limit of local window. Specifically, a similar idea of coupled dictionary learning can be adopted in a token dictionary learning manner. In the following subsection Sec. 3.2, we introduce how we establish a token dictionary to learn typical structures from the training dataset and utilize cross attention operation to provide the learned supplementary information to all the image tokens. Moreover, inspired by the image-specific online dictionary learning approach [27, 28], we further propose an adaptive dictionary refinement strategy in subsection Sec. 3.3. By refining the dictionary with activated tokens, we could adapt the learned external dictionary to image-specific dictionary to better fit the image content and propagate the summarized global information to all the image tokens. Another advantage of the introduced token dictionary lies in its similarity matrix with the image tokens. According to the indexes of the closest dictionary items, we are able to group image tokens into categories. Instead of leveraging image tokens in the same local window to enhance image feature, the proposed category-based self-attention module (subsection Sec. 3.4) allows us to take benefit from similar tokens from the whole image.

3.2. Token Dictionary Cross-Attention

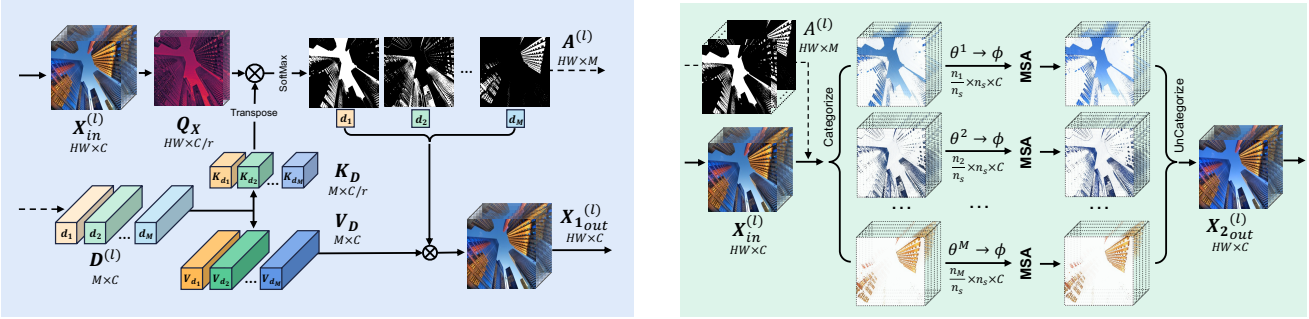
In this subsection, we introduce the details of our proposed token dictionary cross-attention block.

In comparison to the existing multi-head self-attention (MSA), which generates query, key, and value tokens by the input feature itself. We aim to introduce an extra dictionary $\boldsymbol{D} \in \mathbb{R}^{M \times d}$, which is initialized as network parameters, to summarize external priors during the training phase. We use the learned token dictionary $\boldsymbol{D}$ to generate the Key dictionary $\boldsymbol{K}_D$ and the Value dictionary $\boldsymbol{V}_D$ and use the input feature $\boldsymbol{X} \in \mathbb{R}^{N \times d}$ to generate Query tokens:

$$\boldsymbol{Q}_X = \boldsymbol{X}\boldsymbol{W}^Q, \quad \boldsymbol{K}_D = \boldsymbol{D}\boldsymbol{W}^K, \quad \boldsymbol{V}_D = \boldsymbol{D}\boldsymbol{W}^V, \quad (3)$$

where $W^Q \in \mathbb{R}^{d \times d/r}$, $W^K \in \mathbb{R}^{d \times d/r}$ and $W^V \in \mathbb{R}^{d \times d}$ are linear transforms for query tokens, key dictionary tokens and value dictionary tokens, respectively. We set $M \ll N$ to maintain a low computational cost. Meanwhile, the feature dimensions of query tokens and key dictionary tokens are reduced to $1/r$ to decrease model size and complexity, where r is the reduction ratio. Then, we use the key dictionary and the value dictionary to enhance query tokens via cross-

Authorized licensed use limited to: ETH BIBLIOTHEK ZURICH. Downloaded on September 23,2025 at 13:50:29 UTC from IEEE Xplore. Restrictions apply.



(a) Token Dictionary Cross-Attention

(b) Adaptive Category-based Multi-head Self-Attention

Figure 2. The proposed (a) Token Dictionary Cross-Attention (TDCA) and (b) Adaptive Category-based Multi-head Self-Attention (AC-MSA). In Fig. 2b, we omit the details of dividing categories θ into sub-categories ϕ for simplicity and better understanding. More details of TDCA and AC-MSA can be found in Sec. 3.2 and Sec. 3.4.

attention calculation:

$$\begin{aligned} \boldsymbol{A} &= \mathrm{SoftMax}(\mathrm{Sim}_{\cos}(\boldsymbol{Q}_X, \boldsymbol{K}_D)/\tau), \\ \mathrm{TDCA}(\boldsymbol{Q}_X, \boldsymbol{K}_D, \boldsymbol{V}_D) &= \boldsymbol{A}\boldsymbol{V}_D. \end{aligned} \tag{4}$$

In Eq. (4), τ is a learnable parameter for adjusting the range of similarity value; $\mathrm{Sim}_{\cos}(\cdot,\cdot)$ represents calculating cosine similarity between two tokens, and $\boldsymbol{S} = \mathrm{Sim}_{\cos}(\boldsymbol{Q}_X, \boldsymbol{K}_D) \in \mathbb{R}^{N\times M}$ is the similarity map between query image tokens and the key dictionary tokens. We use the normalized cosine distance instead of the dot product operation in MSA because we want each token in the dictionary to have an equal opportunity to be selected, and the similar magnitude normalization operation is commonly used in previous dictionary learning works. Then we use a SoftMax function to transform the similarity map $\boldsymbol{S}$ to attention map $\boldsymbol{A}$ for subsequent calculations.

The above TDCA operation first selects similar tokens in key dictionary and obtains the attention map, which is similar to the sparse representation process in Eq. (1) to obtain representation coefficients; then TDCA utilizes the similarity values to combine the corresponding tokens in value dictionary, which is the same as reconstructing HR patch with HR dictionary atoms and representation coefficients. By this way, our TDCA is able to embed the external prior into the learned dictionary to enhance the input image feature. We will validate the effectiveness of using token dictionary in our ablation study in Sec. 4.2.

3.3. Adaptive Dictionary Refinement

In the previous subsection, we have presented how to incorporate extra token dictionary to supply external prior for super-resolution transformer. Since the image features in each layer are projected to different feature spaces by Multi-Layer Perceptrons (MLPs), we need to learn different Token Dictionary for each layer to provide external prior in each specific feature space. This will result in a large number of additional parameters. In this subsection, we introduce an adaptive refining strategy that refines the token dictionary of the previous layer based on the similarity map and updated features in a reversed form of attention.

To introduce the proposed adaptive refining strategy, we set up the layer index (l) for the input features and token dictionary, i.e. $\boldsymbol{X}^{(l)}$ and $\boldsymbol{D}^{(l)}$ denote input feature and token dictionary of the l-th layer, respectively. We only establish a token dictionary for the initial layer $\boldsymbol{D}^{(1)}$ as network parameter discussed in Sec. 3.2 to incorporate external prior knowledge. In the following layers, each dictionary $\boldsymbol{D}^{(l)}$ is refined based on $\boldsymbol{D}^{(l-1)}$ from the previous layer. For each token in the dictionary $\{\boldsymbol{d}_i^{(l)}\}_{i=1,\dots,M}$ of the l-th layer, we select the corresponding similar tokens in the enhanced feature $\boldsymbol{X}^{(l+1)}$, i.e., the output of the l-th layer to refine it. To be more specific, we denote $\boldsymbol{a}_i^{(l)}$ as the i-th column of attention map $\boldsymbol{A}^{(l)}$, which contains the attention weight between $\boldsymbol{d}_i^{(l)}$ and all the N query tokens $\boldsymbol{X}^{(l)}$. Therefore, based on each $\boldsymbol{a}_i^{(l)}$, we can select the corresponding enhanced tokens $\boldsymbol{X}^{(l+1)}$ to reconstruct the new token dictionary element $\boldsymbol{d}_i^{(l)}$ and combine them to form $\boldsymbol{D}^{(l)}$:

$$\begin{aligned} \hat{\boldsymbol{D}}^{(l)} &= \mathrm{SoftMax}(\mathrm{Norm}({\boldsymbol{A}^{(l)}}^T))\boldsymbol{X}^{(l+1)}, \\ \boldsymbol{D}^{(l+1)} &= \sigma\hat{\boldsymbol{D}}^{(l)} + (1-\sigma)\boldsymbol{D}^{(l)}, \end{aligned} \tag{5}$$

where Norm is normalization layer to adjust the range of attention map. This refinement can also be perceived as a reverse form of attention in Eq. (4), summarizing the information of updated feature into token dictionary. Then, based on a learnable parameter σ, we adaptively combine $\hat{\boldsymbol{D}}^{(l)}$ and $\boldsymbol{D}^{(l)}$ to obtain $\boldsymbol{D}^{(l+1)}$. In this way, the refined token dictionary is able to integrate both external prior and specific internal prior of the input image.

Due to the linear complexity of the proposed TDCA with the number of image tokens, we do not need to divide the image into windows and $\boldsymbol{X}^{(l)}$ represents all image tokens. Starting from the initial token dictionary $\boldsymbol{D}^{(1)}$, which introduces external prior into the network, our adaptive refinement strategy gradually selects relevant tokens from the entire image to refine the dictionary. The refined dictionary could cross the boundary of self-attention window to summarize the typical local structures of the whole image and

Authorized licensed use limited to: ETH BIBLIOTHEK ZURICH. Downloaded on September 23,2025 at 13:50:29 UTC from IEEE Xplore. Restrictions apply.

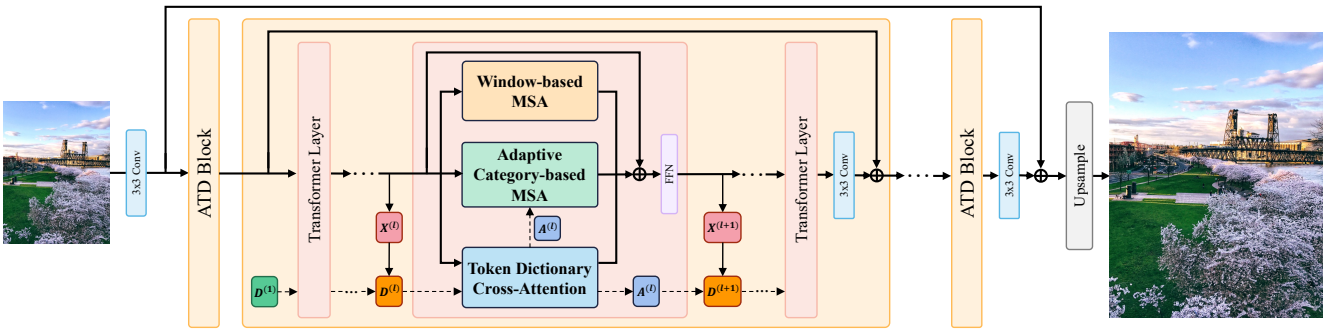


Figure 3. The overall architecture of the proposed ATD network. Token dictionary cross-attention (Fig. 2a), adaptive category-based MSA (Fig. 2b), and window-based MSA [24] form the main structure of the transformer layer. Each ATD block contains several transformer layers and an initial token dictionary $\boldsymbol{D}^{(1)}$. The token dictionary is recurrently adapted via the adaptive dictionary refinement operation.

consequently improve image feature with global information. Furthermore, the class information is implicitly embedded in the refined token dictionary. The attention map $\boldsymbol{A}$ contains the similarity relation between the feature and the token dictionary, which is similar to the image classification task to some extent. The higher similarity between the pixel x_j and a token dictionary atom $\boldsymbol{d}_i$ represents the higher probability that x_j belongs to the class of $\boldsymbol{d}_i$. In the next subsection, we utilize this class information to adaptively partition the input feature and propose a category-based self-attention mechanism to achieve non-local attention while keeping an affordable computational cost.

3.4. Adaptive Category-based Attention

Due to the quadratic computational complexity of self-attention, most of the existing methods, such as Swin Transformer, have to divide the input feature into rectangular windows before performing attention. Such a window-based attention calculation severely limits the scope of receptive field. Furthermore, this content-independent partition strategy could lead to unrelated tokens being grouped into the same window, affecting the accuracy of the attention map.

To make better use of self-attention, adaptive feature partitioning could be an appropriate choice. Thanks to the attention map between input feature and token dictionary obtained by TDCA, which implicitly incorporates the class information of each pixel, we can categorize the input feature. We classify each pixel into various categories $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \cdots, \boldsymbol{\theta}^M$ based on which dictionary token is most similar to:

$$\boldsymbol{\theta}^i = \{x_j | \arg\max_k(\boldsymbol{A}_{jk}) = i\}, \tag{6}$$

where $\boldsymbol{A} \in \mathbb{R}^{N \times M}$ is the attention map obtained by Eq. (4). The pixel x_j will be classified into $\boldsymbol{\theta}^i$ if $\boldsymbol{A}_{ji}$ is the highest among $\boldsymbol{A}_{j1}, \boldsymbol{A}_{j2}, \cdots, \boldsymbol{A}_{jM}$, which indicates that x_j is more likely to be of the same class as the i-th token $\boldsymbol{d}_i$ in the dictionary. Therefore, each category can be perceived as an irregularly shaped window that contains tokens of the same class. An example of categorization visualization is presented in Fig. 5. However, the number of tokens in each category may differ, which results in low parallelism efficiency and significant computational burden. To address the issue of unbalanced categorization, we refer to [32] to further divide the categories $\boldsymbol{\theta}$ into sub-categories $\boldsymbol{\phi}$:

$$\begin{aligned} \boldsymbol{\phi} &= \left[\boldsymbol{\theta}_1^1, \boldsymbol{\theta}_2^1, \cdots, \boldsymbol{\theta}_{n_1}^1, \cdots, \boldsymbol{\theta}_{n_M}^M\right], \\ \boldsymbol{\phi}^j &= \left[\boldsymbol{\phi}_{j*n_s+1}, \boldsymbol{\phi}_{j*n_s+2}, \cdots, \boldsymbol{\phi}_{(j+1)*n_s}\right], \end{aligned} \tag{7}$$

where the category $\boldsymbol{\theta}^i$ contains n_i tokens. Each category is flattened and concatenated to form $\boldsymbol{\phi}$, then divided into sub-categories $[\boldsymbol{\phi}^1, \boldsymbol{\phi}^2, \cdots, \boldsymbol{\phi}^j, \cdots]$. After division, all subcategories have the same fixed size n_s, improving parallelism efficiency. The illustrations are presented in the supplementary. In general, the procedure of AC-MSA can be formulated as:

$$\begin{aligned} \{\boldsymbol{\phi}^j\} &= \text{Categorize}(\boldsymbol{X}_{in}), \\ \hat{\boldsymbol{\phi}}^j &= \text{MSA}(\boldsymbol{\phi}^j\boldsymbol{W}^Q, \boldsymbol{\phi}^j\boldsymbol{W}^K, \boldsymbol{\phi}^j\boldsymbol{W}^V), \\ \boldsymbol{X}_{out} &= \text{UnCategorize}(\{\hat{\boldsymbol{\phi}}^j\}), \end{aligned} \tag{8}$$

where the Categorize operation is the combination of Eq. (6) and Eq. (7) that divides input feature into categories and further into sub-categories $\{\boldsymbol{\phi}^j\}$. Then, we could view each $\boldsymbol{\phi}^j$ as an attention group and perform multi-head self-attention within each group. Finally, we use the UnCategorize operation (inversed Categorize operation) to put each pixel back to its original position on the feature map to form $\boldsymbol{X}_{out}$.

While the subdivision into sub-categories restricts the size of each attention group, its impact on the receptive field is minimal. This is due to the sort operation in Eq. (7) which involves random shuffle. Therefore, each $\boldsymbol{\phi}^j$ can be viewed as a random sample of some category $\boldsymbol{\theta}^i$. The tokens in a certain $\boldsymbol{\phi}^j$ could still be spread throughout the feature map, maintaining a global receptive field. In general, the proposed AC-MSA classifies similar features into the same category and performs attention within each category, breaking through the limitation of window partitioning and establishing global connections between similar features. We will conduct ablation studies and provide visualization of the categorization results in later sections to quantitatively and qualitatively verify the effectiveness of AC-MSA.

Authorized licensed use limited to: ETH BIBLIOTHEK ZURICH. Downloaded on September 23,2025 at 13:50:29 UTC from IEEE Xplore. Restrictions apply.

3.5. The Overall Network Architecture

Having our proposed token dictionary cross-attention (TDCA), adaptive dictionary refinement (ADR) strategy, and adaptive category-based multi-head self-attention (AC-MSA), we are able to establish our Adaptive Token Dictionary (ATD) network for image super-resolution. As shown in Fig. 3, given an input low-resolution image, we first utilize a 3×3 convolution layer to extract shallow features. The shallow features are then fed into a series of ATD blocks, where each ATD block contains several ATD transformer layers. We combine token dictionary cross-attention, adaptive category-based multi-head self-attention, and the commonly used shift window-based multi-head self-attention [22, 24] to form the transformer layer. These three attention modules work in parallel to take advantage of external, global, and local features of the input feature. Then, the features are combined by a summation operation. In addition to the attention module, our transformer layer also utilizes the LayerNorm and FFN layers, which have been commonly utilized in other transformer-based architectures. Moreover, the token dictionary begins with the learnable parameters within each ATD block. It takes part in the token dictionary cross-attention of each transformer layer, and we utilize the adaptive dictionary refinement strategy to adapt the dictionary to the input feature for the next layer. After the ATD blocks, we utilize an extra convolution layer followed by a pixel shuffle operation to generate the final HR estimation.

4. Experiments

4.1. Experimental Settings

We propose the ATD model that employs a sequence of ATD blocks as its backbone. There are six ATD blocks in total, each comprising six transformer layers with a channel number of 210. We establish 128 tokens for our external token dictionary $\boldsymbol{D}^{(1)}$ in each ATD-block and use a reduction rate $r = 10.5$ to decrease the channel number to 20 for similarity calculation. Each dictionary is randomly initialized as a tensor of shape $[128, 210]$ in normal distribution. For the adaptive category-based attention branch, the sub-categories size n_s is set to 128. Furthermore, we establish ATD-light as a lightweight version of ATD with 48 feature dimensions and 4 ATD blocks for lightweight SR task. The number of tokens in each dictionary is reduced to 64, and we also adjust the reduction rate to $r = 6$ to maintain eight dimensions during the similarity calculation. Details of training procedure can be found in the supplementary material.

4.2. Ablation Study

We perform ablation studies on the rescaled ATD-light model and train all models for 250k iterations on the DIV2K [34] dataset. We then evaluate them on Set5 [2], Urban100 [15], and Manga109 [30] benchmarks.

Table 1. Ablation study on the effects of each component. Detailed experimental settings can be found in our Ablation study section.

TDCA	ADR	AC-MSA	Urban100		Manga109	
			PSNR	SSIM	PSNR	SSIM
			26.25	0.7907	30.66	0.9113
√			26.32	0.7929	30.76	0.9118
√	√		26.36	0.7931	30.79	0.9123
√	√	√	26.51	0.7975	30.98	0.9144

Table 2. Ablation study on different designs of category-based attention. CA denotes category-based attention.

Model	Set5		Urban100		Manga109	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
w/o CA	32.30	0.8957	26.25	0.7907	30.66	0.9113
random CA	32.38	0.8962	26.46	0.7955	30.92	0.9139
adaptive CA	32.46	0.8973	26.51	0.7975	30.98	0.9144

Table 3. Ablation study on sub-category size n_s and dictionary size M. The best results are **highlighted**.

n_s	Urban100		Manga109		M	Urban100		Manga109	
	PSNR	SSIM	PSNR	SSIM		PSNR	SSIM	PSNR	SSIM
0	26.36	0.7931	30.79	0.9123	16	26.45	0.7950	30.90	0.9137
64	26.44	0.7948	30.90	0.9131	32	26.49	0.7965	30.97	0.9142
128	26.51	0.7975	30.98	0.9144	64	**26.51**	**0.7975**	**30.98**	**0.9144**
192	**26.55**	**0.7984**	**31.01**	**0.9150**	96	**26.51**	0.7970	30.95	0.9141

Effects of TDCA, ADR, and AC-MSA. In order to show the effectiveness of several key design choices in the proposed adaptive token dictionary (ATD) model, we establish four models and compare their ability for image SR. The first model is the baseline model; we remove the TDCA and AC-MSA branch and only adopt the SW-MSA block to process image features. To demonstrate the effectiveness of learned token dictionary and token dictionary cross-attention, we present the second model, which directly learns an external token dictionary for each Transformer layer. In the third model, we employ the adaptive dictionary refinement strategy to tailor the learned token dictionary to the specific input feature. As shown in Tab. 1, the TDCA branch and the ADR strategy jointly produce 0.11 dB and 0.13 dB improvement on Urban100 and Manga109 datasets respectively. Furthermore, equipped with adaptive category-based MSA, the final model achieves the best performance of 26.51 / 30.98 dB on the Urban100 / Manga109 benchmark. These experimental results clearly demonstrate the advantages of TDCA, ADR, and AC-MSA.

Effects of different designs of category-based attention. We conduct experiments to explore the effectiveness of the category-based partition strategy. First, we evaluate the advantages of category-based attention, using a random token dictionary for rough categorization. The results in Tab. 2 demonstrate that this random category-based attention still performs better than the baseline. Then, with the learned adaptive token dictionary, we can perform the categorization procedure more accurately. The more precise categorization leads to better partition results, resulting in an extra performance gain of 0.05-0.08 dB when using adaptive category-based attention, as opposed to the random one.

Authorized licensed use limited to: ETH BIBLIOTHEK ZURICH. Downloaded on September 23,2025 at 13:50:29 UTC from IEEE Xplore. Restrictions apply.

Table 4. Quantitative comparison (PSNR/SSIM) with state-of-the-art methods on **classical SR** task. The best and second best results are colored with red and blue. Results on $\times 3$ model are presented in the supplementary material.

Method	Scale	Params	Set5		Set14		BSD100		Urban100		Manga109	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR [23]	$\times 2$	42.6M	38.11	0.9602	33.92	0.9195	32.32	0.9013	32.93	0.9351	39.10	0.9773
RCAN [43]	$\times 2$	15.4M	38.27	0.9614	34.12	0.9216	32.41	0.9027	33.34	0.9384	39.44	0.9786
SAN [9]	$\times 2$	15.7M	38.31	0.9620	34.07	0.9213	32.42	0.9028	33.10	0.9370	39.32	0.9792
HAN [33]	$\times 2$	63.6M	38.27	0.9614	34.16	0.9217	32.41	0.9027	33.35	0.9385	39.46	0.9785
IPT [3]	$\times 2$	115M	38.37	-	34.43	-	32.48	-	33.76	-	-	-
SwinIR [22]	$\times 2$	11.8M	38.42	0.9623	34.46	0.9250	32.53	0.9041	33.81	0.9433	39.92	0.9797
EDT [20]	$\times 2$	11.5M	38.45	0.9624	34.57	0.9258	32.52	0.9041	33.80	0.9425	39.93	0.9800
CAT-A [5]	$\times 2$	16.5M	38.51	0.9626	34.78	0.9265	32.59	0.9047	34.26	0.9440	40.10	0.9805
ART [41]	$\times 2$	16.4M	38.56	0.9629	34.59	0.9267	32.58	0.9048	34.30	0.9452	40.24	0.9808
HAT [4]	$\times 2$	20.6M	38.63	0.9630	34.86	0.9274	32.62	0.9053	34.45	0.9466	40.26	0.9809
ATD (ours)	$\times 2$	20.1M	38.61	0.9629	34.92	0.9275	32.64	0.9054	34.73	0.9476	40.35	0.9810
EDSR [23]	$\times 4$	43.0M	32.46	0.8968	28.80	0.7876	27.71	0.7420	26.64	0.8033	31.02	0.9148
RCAN [43]	$\times 4$	15.6M	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087	31.22	0.9173
SAN [9]	$\times 4$	15.9M	32.64	0.9003	28.92	0.7888	27.78	0.7436	26.79	0.8068	31.18	0.9169
HAN [33]	$\times 4$	64.2M	32.64	0.9002	28.90	0.7890	27.80	0.7442	26.85	0.8094	31.42	0.9177
IPT [3]	$\times 4$	116M	32.64	-	29.01	-	27.82	-	27.26	-	-	-
SwinIR [22]	$\times 4$	11.9M	32.92	0.9044	29.09	0.7950	27.92	0.7489	27.45	0.8254	32.03	0.9260
EDT [20]	$\times 4$	11.6M	32.82	0.9031	29.09	0.7939	27.91	0.7483	27.46	0.8246	32.05	0.9254
CAT-A [5]	$\times 4$	16.6M	33.08	0.9052	29.18	0.7960	27.99	0.7510	27.89	0.8339	32.39	0.9285
ART [41]	$\times 4$	16.6M	33.04	0.9051	29.16	0.7958	27.97	0.7510	27.77	0.8321	32.31	0.9283
HAT [4]	$\times 4$	20.8M	33.04	0.9056	29.23	0.7973	28.00	0.7517	27.97	0.8368	32.48	0.9292
ATD (ours)	$\times 4$	20.3M	33.14	0.9061	29.25	0.7976	28.02	0.7524	28.22	0.8414	32.65	0.9308

Table 5. Quantitative comparison (PSNR/SSIM) with state-of-the-art methods on **lightweight SR** task. The best and second best results are colored with red and blue. Results on $\times 3$ model are presented in the supplementary material.

Method	Scale	Params	Set5		Set14		BSD100		Urban100		Manga109	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
CARN [1]	$\times 2$	1,592K	37.76	0.9590	33.52	0.9166	32.09	0.8978	31.92	0.9256	38.36	0.9765
IMDN [16]	$\times 2$	694K	38.00	0.9605	33.63	0.9177	32.19	0.8996	32.17	0.9283	38.88	0.9774
LAPAR-A [19]	$\times 2$	548K	38.01	0.9605	33.62	0.9183	32.19	0.8999	32.10	0.9283	38.67	0.9772
LatticeNet [26]	$\times 2$	756K	38.15	0.9610	33.78	0.9193	32.25	0.9005	32.43	0.9302	-	-
SwinIR-light [22]	$\times 2$	910K	38.14	0.9611	33.86	0.9206	32.31	0.9012	32.76	0.9340	39.12	0.9783
ELAN [42]	$\times 2$	582K	38.17	0.9611	33.94	0.9207	32.30	0.9012	32.76	0.9340	39.11	0.9782
SwinIR-NG [6]	$\times 2$	1181K	38.17	0.9612	33.94	0.9205	32.31	0.9013	32.78	0.9340	39.20	0.9781
OmniSR [36]	$\times 2$	772K	38.22	0.9613	33.98	0.9210	32.36	0.9020	33.05	0.9363	39.28	0.9784
ATD-light (Ours)	$\times 2$	753K	38.29	0.9616	34.10	0.9217	32.39	0.9023	33.27	0.9375	39.52	0.9789
CARN [1]	$\times 4$	1,592K	32.13	0.8937	28.60	0.7806	27.58	0.7349	26.07	0.7837	30.47	0.9084
IMDN [16]	$\times 4$	715K	32.21	0.8948	28.58	0.7811	27.56	0.7353	26.04	0.7838	30.45	0.9075
LAPAR-A [19]	$\times 4$	659K	32.15	0.8944	28.61	0.7818	27.61	0.7366	26.14	0.7871	30.42	0.9074
LatticeNet [26]	$\times 4$	777K	32.30	0.8962	28.68	0.7830	27.62	0.7367	26.25	0.7873	-	-
SwinIR-light [22]	$\times 4$	930K	32.44	0.8976	28.77	0.7858	27.69	0.7406	26.47	0.7980	30.92	0.9151
ELAN [42]	$\times 4$	582K	32.43	0.8975	28.78	0.7858	27.69	0.7406	26.54	0.7982	30.92	0.9150
SwinIR-NG [6]	$\times 4$	1201K	32.44	0.8980	28.83	0.7870	27.73	0.7418	26.61	0.8010	31.09	0.9161
OmniSR [36]	$\times 4$	792K	32.49	0.8988	28.78	0.7859	27.71	0.7415	26.65	0.8018	31.02	0.9151
ATD-light (Ours)	$\times 4$	769K	32.63	0.8998	28.89	0.7886	27.79	0.7440	26.97	0.8107	31.48	0.9198

Effects of sub-category size n_s. Increasing the window size is essential for window-based attention. A larger window size provides a wider range of receptive fields, which in turn leads to improved performance. We carry out experiments to explore the influence of varying sub-category sizes n_s from 0 to 192 on AC-MSA, where 0 represents the removal of the category-based branch, as illustrated in Tab. 3. The model is significantly improved when the value of n_s is raised to 128. However, when we continue increasing n_s, the model performance improves slowly. This is because AC-MSA has the ability to model long-range dependencies with an appropriate sub-category size. The larger n_s contributes less to the receptive field and reconstruction accuracy. To balance performance and computational resource consumption, we set $n_s = 128$ for our final model.

Effects of token dictionary size M. In the token dictionary cross-attention branch, we initialize the token dictionary as M learnable vectors. We investigate the performance change by gradually increasing the dictionary size from 16 to 96. As shown in Tab. 3, increasing the dictionary size yields an improvement of $0.06 - 0.08$ dB on the evaluation benchmark at first. However, when M is set to 96, the model even shows performance degradation. It indicates that the excess of tokens exceeds the modeling capacity of the model and results in unsatisfactory outcomes.

4.3. Comparisons with State-of-the-Art Methods

We choose the commonly used Set5 [2], Set14 [40], BSD100 [29], Urban100 [15], and Manga109 [30] as evaluation datasets and compare the proposed ATD model with current state-of-the-art SR methods.

We first compare our method with the state-of-the-art classical SR methods: EDSR [23], RCAN [43], SAN [9], HAN [33], IPT [3], EDT [20], SwinIR [22], CAT [5], ART [41], HAT [4]. The results are presented in Tab. 4 With comparable parameter size, the proposed ATD model

Authorized licensed use limited to: ETH BIBLIOTHEK ZURICH. Downloaded on September 23,2025 at 13:50:29 UTC from IEEE Xplore. Restrictions apply.

Figure 4. Visual comparisons of **ATD** and other state-of-the-art image super-resolution methods.

(a) (b) (c) (d) (e)

Figure 5. Visualization of categorization results of adaptive category-based MSA. (a) is the input image. The white part of each binarized image from (b) - (e) represents a single attention category.

significantly outperforms HAT [4]. Specifically, the ATD yields $0.25 - 0.28$ dB PSNR gains on the Urban100 dataset for different zooming factors. For the lightweight SR task, we compare our method with CARN [1], IMDN [16], LAPAR [19], LatticeNet [26], SwinIR [22], SwinIR-NG [6], ELAN [42], and OmniSR [36]. As shown in Tab. 5, the proposed ATD-light consistently outperforms recent lightweight method OmniSR [36] on all benchmark datasets. Our ATD-light surpasses OmniSR by a large margin (0.46dB) on the $\times 4$ Manga109 benchmark. Equipped with the token dictionary and category-based attention, our ATD-light model could make better use of external prior to recover HR details under challenging conditions.

We also provide some visual examples using different methods to qualitatively verify the efficacy of ATD, as shown in Fig. 4. These images clearly demonstrate our advantage in recovering sharp edges and clean textures from severely degraded LR input. More visual examples can be found in the supplementary material.

4.4. Visualization Analysis

We further visualize the categorization results in Fig. 5 to verify the effectiveness of the category-based attention mechanism. We use the binarized images to symbolize each attention category. These illustrations clearly show that visually or semantically similar pixels are grouped together. Specifically, most of trees and shrubs are grouped in (b) and (c); the roof part is classified into (d), and (e) is dominated by the area of smooth texture in the image. It indicates that the external prior knowledge of class information is incorporated into the token dictionary. Therefore, AC-MSA can classify similar features into the same attention category, improving the accuracy of the attention map and performance. This again confirms the rationality and effectiveness of category-based attention mechanism.

5. Conclusion

In this paper, we proposed a new Transformer-based super-resolution network. Inspired by traditional dictionary learning methods, we proposed learning token dictionaries to provide external supplementary information to estimate the missing high-quality details. We then proposed an adaptive dictionary refinement strategy which could utilize the similarity map of the preceding layer to refine the learned dictionary, allowing it to better fit the content of a specific input image. Furthermore, with the external prior embedding in the token dictionary, we proposed to categorize input features and perform self-attention within each category. This category-based attention transcends the limit of local window, establishing long-range connections between similar structures across the image. We conducted ablation studies to demonstrate the effectiveness of the proposed token dictionary, adaptive refinement strategy, and category-based attention. The extensive experimental results on multiple benchmark datasets illustrate that our method has achieved state-of-the-art results on single-image super-resolution.

Authorized licensed use limited to: ETH BIBLIOTHEK ZURICH. Downloaded on September 23,2025 at 13:50:29 UTC from IEEE Xplore. Restrictions apply.

References

[1] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network, 2018. 7, 8

[2] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie-line Alberi Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Procedings of the British Machine Vision Conference 2012*, 2012. 6, 7

[3] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer, 2020. 2, 7

[4] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22367–22377, 2023. 1, 2, 7, 8

[5] Zheng Chen, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Cross aggregation transformer for image restoration. In *NeurIPS*, 2022. 2, 7

[6] Haram Choi, Jeongmin Lee, and Jihoon Yang. N-gram in swin transformers for efficient lightweight image super-resolution, 2022. 7, 8

[7] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers, 2021. 2

[8] Marcos V Conde, Ui-Jin Choi, Maxime Burchi, and Radu Timofte. Swin2sr: Swinv2 transformer for compressed image super-resolution and restoration. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 669–687. Springer, 2023. 3

[9] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 7

[10] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 295–307, 2015. 1, 2

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. 2

[12] Shuhang Gu, Wangmeng Zuo, Qi Xie, Deyu Meng, Xiangchu Feng, and Lei Zhang. Convolutional sparse coding for image super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1823–1831, 2015. 3

[13] Shuhang Gu, Shi Guo, Wangmeng Zuo, Yunjin Chen, Radu Timofte, Luc Van Gool, and Lei Zhang. Learned dynamic guidance for depth image reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2437–2452, 2019. 2

[14] He He and Wan-Chi Siu. Single image super-resolution using gaussian process regression. In *CVPR 2011*, pages 449–456. IEEE, 2011. 1

[15] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 6, 7

[16] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *Proceedings of the 27th ACM International Conference on Multimedia*, 2019. 7, 8

[17] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2

[18] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1637–1645, 2016. 2

[19] Wenbo Li, Kun Zhou, Lu Qi, Nianjuan Jiang, Jiangbo Lu, and Jiaya Jia. Lapar: Linearly-assembled pixel-adaptive regression network for single image super-resolution and beyond, 2020. 7, 8

[20] Wenbo Li, Xin Lu, Jiangbo Lu, Xiangyu Zhang, and Jiaya Jia. On efficient transformer and image pre-training for low-level vision. *arXiv preprint arXiv:2112.10175*, 2021. 2, 3, 7

[21] Yawei Li, Yuchen Fan, Xiaoyu Xiang, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Efficient and explicit modelling of image hierarchies for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2, 3

[22] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer, 2021. 1, 2, 3, 6, 7, 8

[23] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017. 1, 2, 7

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. 2, 3, 5, 6

[25] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 3

[26] Xiaotong Luo, Yuan Xie, Yulun Zhang, Yanyun Qu, Cuihua Li, and Yun Fu. *LatticeNet: Towards Lightweight Image Super-Resolution with Lattice Block*, page 272–289. 2020. 7, 8

[27] Julien Mairal, Jean Ponce, Guillermo Sapiro, Andrew Zisserman, and Francis Bach. Supervised dictionary learning. *Advances in neural information processing systems*, 21, 2008. 3

[28] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse

Authorized licensed use limited to: ETH BIBLIOTHEK ZURICH. Downloaded on September 23,2025 at 13:50:29 UTC from IEEE Xplore. Restrictions apply.

coding. *Journal of Machine Learning Research*, 11(1), 2010. 3

[29] 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*, 2002. 7

[30] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, page 21811–21838, 2016. 6, 7

[31] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S Huang, and Humphrey Shi. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[32] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 5

[33] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. *Single Image Super-Resolution via a Holistic Attention Network*, page 191–207. 2020. 2, 7

[34] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, Lei Zhang, Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, Kyoung Mu Lee, et al. Ntire 2017 challenge on single image super-resolution: Methods and results. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017. 6

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[36] Hang Wang, Xuanhong Chen, Bingbing Ni, Yutian Liu, and Liu jinfan. Omni aggregation networks for lightweight image super-resolution. In *Conference on Computer Vision and Pattern Recognition*, 2023. 2, 7, 8

[37] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2022. 2

[38] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 2

[39] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010. 1, 3

[40] Roman Zeyde, Michael Elad, and Matan Protter. *On Single Image Scale-Up Using Sparse-Representations*, page 711–730. 2012. 3, 7

[41] Jiale Zhang, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Accurate image restoration with attention retractable transformer. In *ICLR*, 2023. 2, 7

[42] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. In *European Conference on Computer Vision*, 2022. 2, 7, 8

[43] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. *Image Super-Resolution Using Very Deep Residual Channel Attention Networks*, page 294–310. 2018. 1, 2, 7

[44] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution, 2018. 2

Authorized licensed use limited to: ETH BIBLIOTHEK ZURICH. Downloaded on September 23,2025 at 13:50:29 UTC from IEEE Xplore. Restrictions apply.