Efficient Attention-Sharing Information Distillation Transformer for Lightweight Single Image Super-Resolution

Karam Park[1], Jae Woong Soh[2*], Nam Ik Cho[1,3*]

[1]Department of ECE, INMC, Seoul National University, Seoul, Korea
[2]Department of EECS, Gwangju Institute of Science and Technology, Gwangju, Korea
[3]Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul, Korea
saturnian77@ispl.snu.ac.kr, jaewoongsoh@gist.ac.kr, nicho@snu.ac.kr

Abstract

Transformer-based Super-Resolution (SR) methods have demonstrated superior performance compared to convolutional neural network (CNN)-based SR approaches due to their capability to capture long-range dependencies. However, their high computational complexity necessitates the development of lightweight approaches for practical use. To address this challenge, we propose the Attention-Sharing Information Distillation (ASID) network, a lightweight SR network that integrates attention-sharing and an information distillation structure specifically designed for Transformer-based SR methods. We modify the information distillation scheme, originally designed for efficient CNN operations, to reduce the computational load of stacked self-attention layers, effectively addressing the efficiency bottleneck. Additionally, we introduce attention-sharing across blocks to further minimize the computational cost of self-attention operations. By combining these strategies, ASID achieves competitive performance with existing SR methods while requiring only around 300 K parameters – significantly fewer than existing CNN-based and Transformer-based SR models. Furthermore, ASID outperforms state-of-the-art SR methods when the number of parameters is matched, demonstrating its efficiency and effectiveness. The code and supplementary material are available on the project page.

Project Page — https://github.com/saturnian77/ASID

Introduction

Single Image Super-Resolution (SISR) is a technology designed to improve the resolution and quality of low-resolution (LR) images by transforming them into high-resolution (HR) counterparts. SISR is widely applied in tasks that require HR imagery, including medical imaging (Shi et al. 2013), satellite imaging (Thornton, Atkinson, and Holland 2006), and surveillance (Zhang et al. 2010). Despite its widespread use, SISR remains a challenging problem due to its inherent ill-posed nature.

Recently, SISR methods based on convolutional neural networks (CNNs) have shown significant performance improvement over traditional approaches (Timofte, De Smet, and Van Gool 2013; Yang et al. 2010; Kim and Kwon 2010).
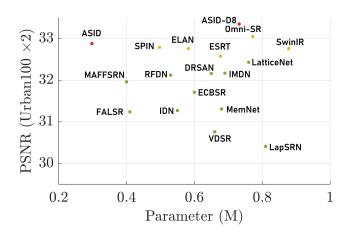
Figure 1: Visualized comparison of PSNR and the number of parameters on the Urban100 ($\times 2$) dataset. Our ASID is compared with state-of-the-art lightweight SR methods. Green markers represent CNN-based methods, while yellow markers denote Transformer-based methods.

Researchers have progressively deepened the CNNs to improve performance using various techniques such as pixel-shuffle operation (Shi et al. 2016) and residual connections (He et al. 2016). However, increasing network depth has resulted in a substantial rise in computational costs. Additionally, the introduction of attention mechanisms (Zhang et al. 2018; Dai et al. 2019; Mei et al. 2020) has further escalated computational demands, limiting the practicality of SR models. As a result, there is growing interest in developing efficient, lightweight networks that balance resource constraints and reconstruction performance.

To address the computational burden of CNNs, various strategies have been introduced, such as weight-sharing (Kim, Kwon Lee, and Mu Lee 2016; Tai, Yang, and Liu 2017; Tai et al. 2017; Li et al. 2019) and feature reuse (He et al. 2019; Luo et al. 2020; Park, Soh, and Cho 2021). However, these approaches often impose limitations on network capabilities and may compromise performance due to restricted depth and receptive fields inherent in convolution operations.

Recently, Transformers (Vaswani et al. 2017) are increasingly being adopted for image reconstruction tasks, in-
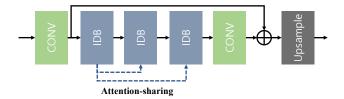
Figure 2: Visualized overall structure of ASID. ASID mainly consists of a convolutional layer for shallow feature extraction, a series of IDBs, and an upsampling module that reconstructs features into an SR image. Blue arrows indicate the attention-sharing mechanism.
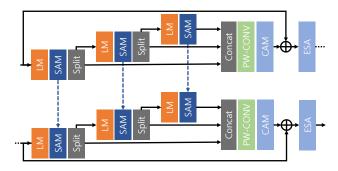

Figure 3: Visualized structure of Information Distillation Blocks (IDBs) and the attention-sharing mechanism. Blue arrows represent the attention-sharing mechanism. PW-CONV denotes pixel-wise convolution.

cluding SISR. Transformer-based SR methods have outperformed CNN-based methods by leveraging non-local context and globally expanding the receptive field. However, the computational intensity of self-attention layers in Transformers poses significant challenges for practical application in SISR. To address this, Liang et al. (2021) proposed a patch-based processing technique to reduce computational demands, though this method limits the receptive field to the local window size. On the other hand, Zamir et al. (2022) introduced self-attention layers that utilize channel correlations to reduce computational costs, yet this approach fails to capture spatial correlations. While these methods represent notable progress in alleviating computational burdens, further advancements are required to improve reconstruction performance and efficiency.

To address the challenges in designing efficient SR Transformers, we introduce a lightweight design scheme called the Attention-Sharing Information Distillation (ASID) network. Unlike previous approaches such as ESRT (Lu et al. 2022), which adapted lightweight CNN frameworks to Transformers with limited success, ASID leverages an understanding of the distinct roles of convolution and self-attention layers (Park and Kim 2022) by integrating them into its architecture. By designing these layers to complement each other and adapting the information distillation scheme (Hui, Wang, and Gao 2018; Hui et al. 2019; Liu, Tang, and Wu 2020) for Transformer architectures, ASID achieves state-of-the-art performance with remarkably few parameters. Additionally, we incorporate an attention-sharing method and eliminate correlation matrix calculations, significantly reducing computational costs without compromising reconstruction quality. With only about 300 K parameters, ASID not only competes effectively with traditional CNN-based and Transformer-based SR methods but also surpasses state-of-the-art SISR networks of comparable complexities when scaled up.

Our contributions can be summarized in three-folds:

- We introduce a new lightweight SR model that combines an information distillation design scheme with a Transformer framework.

- We propose a feature distillation framework with attention sharing to alleviate the efficiency bottleneck of a self-attention layer, significantly reducing the computational load required for self-attention matrix calculations.

- Our experimental results demonstrate that the proposed network outperforms state-of-the-art methods.

Related Works

CNN-based SISR

There have been many *lightweight networks* for SISR that aimed to find a balance between computational complexity with performance. Early strategies focused on reusing kernel weights (Kim, Kwon Lee, and Mu Lee 2016; Tai, Yang, and Liu 2017; Tai et al. 2017; Li et al. 2019), which effectively reduced the number of parameters but failed to lower computational costs. Additionally, their performance was somewhat limited due to the repetitive use of the same kernels across multiple layers.

To address these shortcomings, researchers also designed efficient and compact network structures. For instance, LatticeNet (Luo et al. 2020) proposed a residual block architecture that exploits multiple potential residual paths. DRSAN (Park, Soh, and Cho 2021) introduced dynamic residual connections that adaptively adjust the residual paths based on the input. The information distillation technique (Hui, Wang, and Gao 2018), which progressively refines features through the information distillation framework, facilitates efficient computation with low parameters and computational loads. This method has been extensively adopted across various lightweight network designs due to its effectiveness and simplicity (Hui et al. 2019; Liu, Tang, and Wu 2020; Kong et al. 2022).

These lightweight SR networks typically employ shallower architectures with fewer channels to reduce complexity. To compensate for their limited capacity, they enhance feature representation using techniques such as feature reuse, residual connection, and dense connection.

Transformer-based SISR

Recently, there have been significant advancements in vision Transformers (Dosovitskiy et al. 2020; Liu et al. 2021; Ranftl, Bochkovskiy, and Koltun 2021), with several attempts made to apply them to image super-resolution (Liang et al. 2021; Zhang et al. 2022; Zamir et al. 2022; Wang et al. 2023). For example, SwinIR (Liang et al. 2021) applied
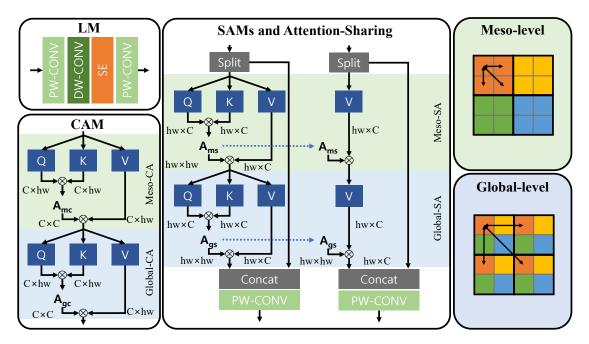
Figure 4: Visualized structure of the Local Module (LM), Spatial Attention Module (SAM), and Channel Attention Module (CAM). The blue arrow represents the attention-sharing mechanism in SAM. By employing the attention-sharing technique, subsequent SAMs bypass the calculation of the spatial attention matrix, which typically accounts for a significant portion of the computational load in self-attention layers. Meso-level self-attention computes attention matrices among pixels within the same partition, whereas global-level self-attention involves pixels from different partitions (Wang et al. 2023). Utilizing both methods effectively mitigates the limited receptive field issue associated with window-based self-attention. All feedforward networks are omitted for simplicity in visualization.

the shifted-window framework (Liu et al. 2021) for image restoration and demonstrated its effectiveness. ESRT (Lu et al. 2022) introduced a new lightweight super-resolution model by combining lightweight CNNs with Transformers. ELAN (Zhang et al. 2022) enhances computational efficiency by sharing the self-attention matrix within a block and the weights of the query and key. Omni-SR (Wang et al. 2023) proposed a lightweight Transformer using omni self-attention, which considers spatial and channel self-attention together. SPIN (Zhang et al. 2023) introduced super-pixel clustering into self-attention operations to reduce self-attention computations. Despite numerous efforts to enhance Transformer efficiency, the structural limitation of stacking self-attention layers continues to hinder their practical application.

Proposed Method

In this section, we provide a comprehensive overview of our ASID framework. We start by introducing the overall structure of the ASID network, followed by a detailed explanation of modules.

Overall Structure

Following the previous state-of-the-art (SOTA) method (Wang et al. 2023) as the baseline, we developed a lightweight Transformer structure that utilizes spatial and channel self-attention at both meso- and global-level. As de-

picted in Fig. 2, ASID mainly consists of the convolution layer for shallow feature extraction, repetitive Information Distillation Blocks (IDBs) for deep feature extraction, and Upsampler for image reconstruction. Given the input image as I_{LR}, the 3×3 convolution layer H_s extracts shallow feature F_0 as

$$F_0 = H_s(I_{LR}). \qquad (1)$$

This convolutional layer maps the input image from the RGB channel to a multi-channel latent feature dimension. Next, cascaded IDBs and the 3×3 convolution layer H_d extract deep features F_d from shallow features F_s as

$$
\begin{aligned}
F_1, A &= H_{IDB_1}(F_0), \\
F_i &= H_{IDB_i}(F_{i-1}, A), \quad i = 2, ...N, \\
F_d &= H_d(F_N),
\end{aligned}
\qquad (2)
$$

where H_{IDB_i} refers to the i-th IDB, A means attention matrices for attention-sharing, and N represents the number of IDB. Each IDB consists of multiple self-attention layers, which extract meaningful features by considering long-range dependencies. These stacked IDBs progressively refine input features, with a final convolution layer H_d used to extract deep features. Then, Upsampler module H_{up} reconstructs high resolution output image I_{SR} from deep features F_d as

$$I_{SR} = H_{up}(F_0 + F_d). \qquad (3)$$

During this process, the pixel-shuffle layer transforms the latent feature dimension back into the RGB channel and HR

space, restoring the output image I_{SR}.

Information Distillation Block

Information distillation processes features by hierarchically splitting them during distillation steps. This method preserves essential representations in one part while passing the remaining features to subsequent modules, recognizing that certain feature channels carry more critical information than others (Zhang et al. 2018; Wang et al. 2021). To adapt this approach for Transformer-based SR, we focus on the fundamental structure of the SR Transformer, which can generally be divided into a local feature extraction module and a self-attention calculation module. We design the Information Distillation Block (IDB) for Transformers by aligning the local feature extraction module with convolutional layers for extracting local features and the self-attention modules with sequential calculation units.

As shown in Fig. 3, IDB refines features in a progressive manner with split and concatenation. IDB is mainly composed of Local Module (LM), Spatial Attention Module (SAM), and Channel Attention Module (CAM). Given the input feature F_{in}, $1st$ IDB refines features as

$$
\begin{aligned}
F_1, A_1 &= SAM_1(LM_1(F_{in})), \\
F_1^{refined}, F_1^{coarse} &= Split(F_1), \\
F_2, A_2 &= SAM_2(LM_2(F_1^{coarse})), \\
F_2^{refined}, F_2^{coarse} &= Split(F_2), \\
F_3, A_3 &= SAM_3(LM_3(F_2^{coarse})),
\end{aligned} \quad (4)
$$

where LM_i represents the i-th LM, SAM_i means the i-th SAM, A_i signifies the i-th spatial self-attention matrix, and $Split$ refers to the channel split. Like the information distillation scheme, the IDB divides features by channels and progressively refines them. LM applies convolution operations to the input features, extracting local features for subsequent self-attention modules. Next, SAM enhances features by capturing pixel correlations both within a single window (meso-level) and across different windows (global-level). After progressive feature refinement, refined hierarchical features are aggregated through concatenation and processed by the 1×1 convolution layer as

$$
\begin{aligned}
F^{refined} &= Conv_{1\times1}(Concat(F_1^{refined}, F_2^{refined}, F_3), \\
F_{out} &= ESA(CAM(F^{refined}) + F_{in}),
\end{aligned} \quad (5)
$$

where CAM represents CAM module, $Conv_{1\times1}$ refers to pixel-wise convolution, and ESA means enhanced spatial-attention operation (Liu et al. 2020). CAM calculates the affinity matrix for intra-window and inter-window pixels across the channel dimension, similar to the operation of SAM. ESA module employs strided convolution and max pooling to achieve a large receptive field and rescale features, thereby enhancing them in a spatial context.

Meanwhile, subsequent IDBs utilize attention-sharing to efficiently perform self-attention layer operations using the spatial self-attention matrices computed by the initial IDB.

This can be represented as:

$$
\begin{aligned}
F_1^{refined}, F_1^{coarse} &= Split(SAM_1(LM_1(F), A_1)), \\
F_2^{refined}, F_2^{coarse} &= Split(SAM_2(LM_2(F_1^{coarse}), A_2)), \\
F_3 &= SAM_3(LM_3(F_2^{coarse}), A_3), \\
F^{refined} &= Conv_{1\times1}(Concat(F_1^{refined}, F_2^{refined}, F_3), \\
F_{out} &= ESA(CAM(F^{refined}) + F).
\end{aligned} \quad (6)
$$

These subsequent IDBs skip the self-attention matrix computation by sharing the initial IDB's attention matrix.

Local Module   We use LM shown in Fig. 4 to extract local features for self-attention layers. LM is composed of pixel-wise convolution, depth-wise convolution, and a squeeze-and-excitation (Hu, Shen, and Sun 2018) module. As shown in Fig. 3, LM is positioned in front of SAM so that SAM can utilize local information to calculate the spatial correlation matrix.

Spatial and Channel Attention Module   The proposed SAM and CAM employ a window-based self-attention method that processes input features into non-overlapping patches. Both modules operate in two stages, referred to as meso-level and global-level (Wang et al. 2023). As shown in Fig. 4, the meso-level self-attention layer calculates the affinity matrix for pixels within individual patches, while the global-level self-attention layer computes the affinity matrix for pixels across different patches. By incorporating information both within and between patches, the proposed method enhances the network's representational capacity, effectively capturing both local information and long-range dependencies.

Firstly, the input feature f is partitioned into windows for both meso-level and global-level attention. For meso-level, the input feature is partitioned into $P \times P$ non-overlapping patches: $(HW \times C) \rightarrow ((h \times P)(w \times P) \times C) \rightarrow (hw \times P^2 \times C)$, where P represents the partition size of meso-level self-attention, and $\rightarrow$ means reshape operation. For global-level, the input feature is partitioned into $G \times G$ non-overlapping patches: $(HW \times C) \rightarrow ((G \times h)(G \times w) \times C) \rightarrow (hw \times G^2 \times C)$, where G represents the partition size of global-level self-attention.

Then, a linear projection is applied to the input feature to compute the self-attention matrix. Given the input feature f, the query f_q, key f_k, and value f_v are calculated as

$$
f_q = Q(f), \quad f_k = K(f), \quad f_v = V(f), \quad (7)
$$

where Q, K, V represent linear projection operation. Next, self-attention matrix A is calculated as

$$
\begin{aligned}
A &= SoftMax(f_q f_k^T), \\
f_{out} &= FFN(Af_v + f),
\end{aligned} \quad (8)
$$

where f_{out} represents the output of self-attention layer, and FFN refers to feedforward network.

Attention-Sharing and Channel-Split   The computational burden of self-attention is a key inefficiency in SR Transformers, yet stacking spatial self-attention layers is

Scale	Methods	Params	Set5 PSNR/SSIM	Set14 PSNR/SSIM	B100 PSNR/SSIM	Urban100 PSNR/SSIM
×2	**ASID**[†]	**298K**	**38.17/0.9613**	**33.96/0.9216**	**32.31/0.9017**	**32.88/0.9353**
	MAFFSRN	402K	37.97/0.9603	33.49/0.9170	32.14/0.8994	31.96/0.9268
	SPIN[†]	497K	38.20/0.9615	33.90/0.9215	32.31/0.9015	32.79/0.9340
	RFDN	534K	38.05/0.9606	33.68/0.9184	32.16/0.8994	32.12/0.9278
	ELAN[†]	582K	38.17/0.9611	33.94/0.9207	32.30/0.9012	32.76/0.9340
	ESRT[†]	677K	38.03/0.9600	33.75/0.9184	32.25/0.9001	32.58/0.9318
	DRSAN	690K	38.11/0.9609	33.64/0.9185	32.21/0.9005	32.35/0.9304
	IMDN	694K	38.00/0.9605	33.63/0.9177	32.19/0.8996	32.17/0.9283
	ASID-D8[†]	**732K**	**38.32/0.9618**	**34.24/0.9232**	**32.40/0.9028**	**33.35/0.9387**
	LatticeNet	756K	38.15/0.9610	33.78/0.9193	32.25/0.9005	32.43/0.9302
	Omni-SR[†]	772K	38.22/0.9613	33.98/0.9210	32.36/0.9020	33.05/0.9363
	SwinIR[†]	878K	38.14/0.9611	33.86/0.9206	32.31/0.9012	32.76/0.9340
×3	**ASID**[†]	**304K**	**34.58/0.9290**	**30.53/0.8465**	**29.23/0.8096**	**28.68/0.8633**
	RFDN	541K	34.41/0.9273	30.34/0.8420	29.09/0.8050	28.21/0.8525
	SPIN[†]	569K	34.65/0.9293	30.57/0.8464	29.23/0.8089	28.71/0.8627
	ELAN[†]	590K	34.64/0.9288	30.55/0.8463	29.21/0.8081	28.69/0.8624
	IMDN	703K	34.36/0.9270	30.32/0.8417	29.09/0.8046	28.17/0.8519
	ASID-D8[†]	**739K**	**34.84/0.9307**	**30.66/0.8491**	**29.32/0.8119**	**29.08/0.8706**
	DRSAN	740K	34.50/0.9278	30.39/0.8437	29.13/0.8065	28.35/0.8566
	LatticeNet	765K	34.53/0.9281	30.39/0.8424	29.15/0.8059	28.33/0.8538
	ESRT[†]	770K	34.42/0.9268	30.43/0.8433	29.15/0.8063	28.46/0.8574
	Omni-SR[†]	780K	34.70/0.9294	30.57/0.8469	29.28/0.8094	28.84/0.8656
	MAFFSRN	807K	34.45/0.9277	30.40/0.8432	29.13/0.8061	28.26/0.8552
	SwinIR[†]	886K	34.62/0.9289	30.54/0.8463	29.20/0.8082	28.66/0.8624
×4	**ASID**[†]	**313K**	**32.39/0.8964**	**28.80/0.7865**	**27.70/0.7418**	**26.48/0.7978**
	RFDN	550K	32.24/0.8952	28.61/0.7819	27.57/0.7360	26.11/0.7858
	SPIN[†]	555K	32.48/0.8983	28.80/0.7862	27.70/0.7415	26.55/0.7998
	ELAN[†]	601K	32.43/0.8975	28.78/0.7858	27.69/0.7406	26.54/0.7982
	IMDN	715K	32.21/0.8948	28.58/0.7811	27.56/0.7353	26.04/0.7838
	DRSAN	730K	32.30/0.8954	28.66/0.7838	27.61/0.7381	26.26/0.7920
	ASID-D8[†]	**748K**	**32.57/0.8990**	**28.89/0.7898**	**27.78/0.7449**	**26.89/0.8096**
	ESRT[†]	751K	32.19/0.8947	28.69/0.7833	27.69/0.7379	26.39/0.7962
	LatticeNet	777K	32.30/0.8962	28.68/0.7830	27.62/0.7367	26.25/0.7873
	Omni-SR[†]	792K	32.49/0.8988	28.78/0.7859	27.71/0.7415	26.64/0.8018
	MAFFSRN	830K	32.20/0.8953	26.62/0.7822	27.59/0.7370	26.16/0.7887
	SwinIR[†]	897K	32.44/0.8976	28.77/0.7858	27.69/0.7406	26.47/0.7980

Table 1: Quantitative comparison of previous CNN-based lightweight SR models and the proposed method. The † symbol represents a Transformer-based SR method, and **bold** highlights the proposed method.

Scale	Methods	Params	FLOPs	PSNR
×4	**ASID**	**313K**	**17.2G**	**28.80**
	ELAN	601K	43.2G	28.78
	IMDN	715K	40.9G	28.58
	LatticeNet	777K	43.6G	28.68
	SwinIR	897K	49.6G	28.77
	ASID-D8	**748K**	**40.7G**	**28.89**

Table 2: Comparisons of the computational cost of lightweight SR methods on the Set14 (×4) dataset. FLOPs are evaluated on a 720p (1280 × 720) output image. **Bold** indicates the proposed method.

essential for performance. To address this, we propose attention-sharing and channel-split techniques. As shown in Fig. 4, attention-sharing enables the sharing of atten-

tion matrices between self-attention layers, eliminating the need for spatial self-attention operations. This allows layers that share attention matrices to skip the computation of self-attention matrices, reducing the parameters required for affinity matrix calculation. Channel-split restricts the number of channels involved in spatial attention operations, which helps decrease both the computational load and the number of parameters. The combination of attention-sharing and channel-split reduces the complexity of self-attention, enabling the stacking of more layers with fewer parameters.

Experimental Results

Settings

Following previous methods, the DIV2K dataset (Agustsson and Timofte 2017) is selected for training, which contains 800 HR images. During training, 64×64-sized RGB patches are utilized, with random flips and rotations applied
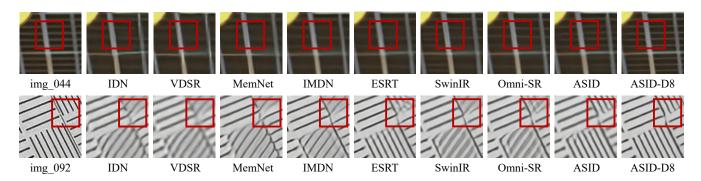
Figure 5: Qualitative Comparison of previous CNN-based and Transformer-based SR methods on the Urban100 (×4) dataset. Note that ASID accurately restores images while using an extremely small number of model parameters.

for data augmentation. For evaluation, we use well-known benchmark datasets, including Set5 (Bevilacqua et al. 2012), Set14 (Zeyde, Elad, and Protter 2010), B100 (Martin et al. 2001), and Urban100 (Huang, Singh, and Ahuja 2015). The performance of the network is evaluated using PSNR and SSIM on the Y channel. For training, we use the ADAM optimizer with a mini-batch size of 16. All models in quantitative comparisons are trained for 1000 epochs, while the models used in ablation studies are trained for 200 epochs. The initial learning rate is set to 5×10^{-4} and is halved every 250 epochs.

We set the number of IDBs in ASID to 3 and the number of channels to 48. Also, we set the number of calculation units in each IDB to 3, which include LM, SAM, and channel-split layers as shown in Fig. 3. In the channel-split layers within or following SAM, we allocate 12 channels to the refined features, while the remaining channels are designated as coarse features and passed on to the next calculation unit. For the self-attention layers, we set the partition size for meso-level attention to 8 and the grid size for global-level attention to 8.

Comparisons with SOTA Methods

In this section, we compare our ASID with state-of-the-art (SOTA) lightweight CNN-based and Transformer-based SR methods to demonstrate the proposed method's effectiveness. We report quantitative and qualitative comparisons with previous SR methods, including DRSAN (Park, Soh, and Cho 2021), ECBSR (Zhang, Zeng, and Zhang 2021), ELAN (Zhang et al. 2022), ESRT (Lu et al. 2022), FALSR (Chu et al. 2021), IDN (Hui, Wang, and Gao 2018), IMDN (Hui et al. 2019), LapSRN (Lai et al. 2018), LatticeNet (Luo et al. 2020), MAFFSRN (Muqeet et al. 2020), MemNet (Tai et al. 2017), Omni-SR (Wang et al. 2023), RFDN (Liu, Tang, and Wu 2020), SPIN (Zhang et al. 2023), SwinIR (Liang et al. 2021), and VDSR (Kim, Lee, and Lee 2016).

We provide quantitative comparisons between proposed methods and previous lightweight SR methods in Fig. 1 and Table 1. Despite having fewer model parameters, our method significantly outperforms CNN-based SR methods, as shown in Table 1. These results suggest that while CNN-based SR methods mainly process local features, ASID utilizes self-attention to capture long-range dependencies. This

Methods	Params	FLOPs	PSNR
Baseline	**288K**	**62.323G**	32.43
+ ID	401K	87.344G	**32.70**
+ ID & AS	365K	79.119G	32.62
+ ID & CS	316K	67.836G	32.53
+ ID & AS & CS	298K	63.592G	32.53

Table 3: Ablation Studies on the proposed ASID network structure. All results are evaluated on the Urban100 (×2) dataset. ID refers to the Information Distillation framework, AS denotes attention-sharing, and CS represents channel-split.

leads to a broader receptive field, improving image reconstruction accuracy. Moreover, ASID achieves performance comparable to previous Transformer-based SR methods, using less than 60 % of their model parameters. This demonstrates that ASID effectively preserves essential information about long-range dependencies, even with lightweight strategies, maintaining performance on par with more complex methods. We further introduce the ASID-D8 model as an additional option, providing an alternative trade-off between performance and complexity. As shown in Table 2, our approach not only reduces model parameters but also significantly lowers computational costs compared to prior methods. In Fig. 5, we compare zoomed-in results from the Urban100 (×4) dataset, highlighting the superior quality of ASID over previous lightweight methods. More qualitative comparisons and experiments on RealSR dataset (Cai et al. 2019) are provided in the supplementary material.

Ablation Studies

Information Distillation Block We propose lightweight network approaches for ASID, including information distillation structure, attention-sharing, and channel-split. To investigate the effectiveness of these proposed methods, we compare the network performance with and without these elements. For the comparison, we introduce a baseline structure consisting of serially connected LM, SAM, and CAM. The results of the experiments are summarized in Table 3.

First, we observe that the baseline performance is the lowest among all configurations. Subsequently, applying the in-

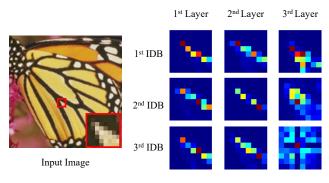1st Layer 2nd Layer 3rd Layer

1st IDB

2nd IDB

3rd IDB

Input Image

Figure 6: Visualized comparison of attention matrices. The attention matrices are collected from models without attention-sharing, as described in Table 3. The visualization depicts the meso-level spatial correlation between the center point and the pixels within the same window.

Scale	Methods	Params	FLOPs	PSNR
×2	IntraGroup	305K	65.317G	32.49
	InterGroup	298K	63.592G	32.53

Table 4: Ablation studies on attention-sharing methods. IntraGroup refers to attention-sharing between adjacent layers within the same block, while InterGroup denotes attention-sharing across blocks, as proposed in our method. PSNR results are evaluated on the Urban100 (×2) dataset.

formation distillation scheme to the baseline results in an increase in performance, which is primarily attributed to the doubling of the number of self-attention layers. However, this increase in self-attention layers also leads to a 40 % rise in both model parameters and computational cost.

Next, we observe that implementing attention-sharing reduces the model parameters and computational load by 10 % each. With the channel-split method, we achieve a 20 % reduction in parameters and computational cost individually. By integrating both methods, the network achieves enhanced performance while maintaining a complexity level that is nearly identical to the baseline. This underscores the effectiveness of these lightweight strategies in enhancing the network's overall efficiency.

Attention-Sharing To investigate suitable attention-sharing methods for our proposed structure, we visualize the spatial attention from an ablation model without attention-sharing in Table 3, specifically focusing on the spatial correlations around the center point for easier visualization. As Fig. 6 shows, even within the same block, the use of spatial attention matrices varies with depth. Therefore, we chose the sharing of spatial attention matrices across blocks of the same depth rather than within the same block. This approach enables the IDB to enhance features by considering various spatial correlations.

Next, we evaluate the effectiveness of the proposed method by comparing two candidate methods for attention-sharing. Table 4 shows two different attention matrix-sharing methods: IntraGroup and InterGroup. IntraGroup
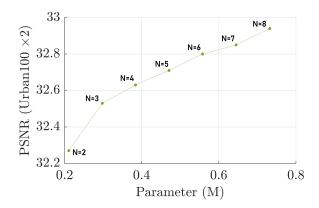

Figure 7: Ablation studies on network depth. Results are evaluated on the Urban100 (×2) dataset. N indicates the number of IDBs implemented in the network.

shares attention matrices within building blocks, while InterGroup shares attention matrices between building blocks. The major difference between the two methods is that IntraGroup enforces adjacent self-attention layers to use the same attention matrix, whereas InterGroup allows adjacent self-attention layers to use different attention matrices. As shown in Table 4, our proposed method shows better results with fewer parameters and FLOPs. This suggests that allowing adjacent self-attention layers to capture diverse pixel correlations improves feature representation.

Network Depth We examine the effect of network depth on performance by adjusting the number of IDBs N from 2 to 8. As shown in Fig. 7, the network's performance consistently improves with an increasing number of blocks, even though correlation matrix computations are omitted due to attention-sharing from the second IDB onward. This demonstrates that the proposed attention-sharing method effectively supports both shallow and deeper network architectures. Considering the trade-off between performance improvements and network complexity, we determine the optimal depth for ASID to be $N = 3$, where performance gains start to diminish, allowing us to retain minimal complexity within the lightweight ASID framework.

Conclusion

We propose the attention-sharing information distillation (ASID) network, a novel lightweight Transformer-based SR method that delivers competitive performance compared to existing lightweight SR methods while utilizing significantly fewer model parameters. ASID employs an information distillation structure specifically adapted for Transformers, enabling the efficient stacking of multiple self-attention layers with low complexity. Additionally, ASID incorporates attention-sharing and channel-split techniques to significantly reduce the computational overhead typically associated with self-attention operations. Experimental results demonstrate that ASID effectively balances model complexity with performance, surpassing previous lightweight SR methods.

Acknowledgments

References

Agustsson, E.; and Timofte, R. 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 126–135.

Bevilacqua, M.; Roumy, A.; Guillemot, C.; and Alberi-Morel, M. L. 2012. Low-complexity single-image super-resolution based on nonnegative neighbor embedding.

Cai, J.; Zeng, H.; Yong, H.; Cao, Z.; and Zhang, L. 2019. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3086–3095.

Chu, X.; Zhang, B.; Ma, H.; Xu, R.; and Li, Q. 2021. Fast, accurate and lightweight super-resolution with neural architecture search. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 59–64. IEEE.

Dai, T.; Cai, J.; Zhang, Y.; Xia, S.-T.; and Zhang, L. 2019. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 11065–11074.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, X.; Mo, Z.; Wang, P.; Liu, Y.; Yang, M.; and Cheng, J. 2019. Ode-inspired network design for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1732–1741.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Huang, J.-B.; Singh, A.; and Ahuja, N. 2015. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5197–5206.

Hui, Z.; Gao, X.; Yang, Y.; and Wang, X. 2019. Lightweight image super-resolution with information multi-distillation network. In *Proceedings of the 27th ACM International Conference on Multimedia*, 2024–2032.

Hui, Z.; Wang, X.; and Gao, X. 2018. Fast and accurate single image super-resolution via information distillation network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 723–731.

Kim, J.; Kwon Lee, J.; and Mu Lee, K. 2016. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1637–1645.

Kim, J.; Lee, J. K.; and Lee, K. M. 2016. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR Oral)*.

Kim, K. I.; and Kwon, Y. 2010. Single-image super-resolution using sparse regression and natural image prior. *IEEE transactions on pattern analysis and machine intelligence*, 32(6): 1127–1133.

Kong, F.; Li, M.; Liu, S.; Liu, D.; He, J.; Bai, Y.; Chen, F.; and Fu, L. 2022. Residual local feature network for efficient super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 766–776.

Lai, W.-S.; Huang, J.-B.; Ahuja, N.; and Yang, M.-H. 2018. Fast and accurate image super-resolution with deep laplacian pyramid networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(11): 2599–2613.

Li, Z.; Yang, J.; Liu, Z.; Yang, X.; Jeon, G.; and Wu, W. 2019. Feedback network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3867–3876.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1833–1844.

Liu, J.; Tang, J.; and Wu, G. 2020. Residual feature distillation network for lightweight image super-resolution. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, 41–55. Springer.

Liu, J.; Zhang, W.; Tang, Y.; Tang, J.; and Wu, G. 2020. Residual feature aggregation network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2359–2368.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Lu, Z.; Li, J.; Liu, H.; Huang, C.; Zhang, L.; and Zeng, T. 2022. Transformer for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 457–466.

Luo, X.; Xie, Y.; Zhang, Y.; Qu, Y.; Li, C.; and Fu, Y. 2020. Latticenet: Towards lightweight image super-resolution with lattice block. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, 272–289. Springer.

Martin, D.; Fowlkes, C.; Tal, D.; and Malik, J. 2001. 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*, volume 2, 416–423. IEEE.

Mei, Y.; Fan, Y.; Zhou, Y.; Huang, L.; Huang, T. S.; and Shi, H. 2020. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5690–5699.

Muqeet, A.; Hwang, J.; Yang, S.; Kang, J.; Kim, Y.; and Bae, S.-H. 2020. Multi-attention based ultra lightweight image super-resolution. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, 103–118. Springer.

Park, K.; Soh, J. W.; and Cho, N. I. 2021. Single Image Super-Resolution with Dynamic Residual Connection. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 1–8. IEEE.

Park, N.; and Kim, S. 2022. HOW DO VISION TRANSFORMERS WORK? In *10th International Conference on Learning Representations, ICLR 2022*.

Ranftl, R.; Bochkovskiy, A.; and Koltun, V. 2021. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12179–12188.

Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1874–1883.

Shi, W.; Caballero, J.; Ledig, C.; Zhuang, X.; Bai, W.; Bhatia, K.; de Marvao, A. M. S. M.; Dawes, T.; O'Regan, D.; and Rueckert, D. 2013. Cardiac image super-resolution with global correspondence using multi-atlas patchmatch. In *International conference on medical image computing and computer-assisted intervention*, 9–16. Springer.

Tai, Y.; Yang, J.; and Liu, X. 2017. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3147–3155.

Tai, Y.; Yang, J.; Liu, X.; and Xu, C. 2017. Memnet: A persistent memory network for image restoration. In *Proceedings of the IEEE international conference on computer vision*, 4539–4547.

Thornton, M. W.; Atkinson, P. M.; and Holland, D. 2006. Sub-pixel mapping of rural land cover objects from fine spatial resolution satellite sensor imagery using super-resolution pixel-swapping. *International Journal of Remote Sensing*, 27(3): 473–491.

Timofte, R.; De Smet, V.; and Van Gool, L. 2013. Anchored neighborhood regression for fast example-based super-resolution. In *Proceedings of the IEEE international conference on computer vision*, 1920–1927.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, H.; Chen, X.; Ni, B.; Liu, Y.; and Liu, J. 2023. Omni aggregation networks for lightweight image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22378–22387.

Wang, L.; Dong, X.; Wang, Y.; Ying, X.; Lin, Z.; An, W.; and Guo, Y. 2021. Exploring sparsity in image super-resolution for efficient inference. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4917–4926.

Yang, J.; Wright, J.; Huang, T. S.; and Ma, Y. 2010. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11): 2861–2873.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5728–5739.

Zeyde, R.; Elad, M.; and Protter, M. 2010. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, 711–730. Springer.

Zhang, A.; Ren, W.; Liu, Y.; and Cao, X. 2023. Lightweight Image Super-Resolution with Superpixel Token Interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12728–12737.

Zhang, L.; Zhang, H.; Shen, H.; and Li, P. 2010. A super-resolution reconstruction algorithm for surveillance images. *Signal Processing*, 90(3): 848–859.

Zhang, X.; Zeng, H.; Guo, S.; and Zhang, L. 2022. Efficient long-range attention network for image super-resolution. In *European conference on computer vision*, 649–667. Springer.

Zhang, X.; Zeng, H.; and Zhang, L. 2021. Edge-oriented convolution block for real-time super resolution on mobile devices. In *Proceedings of the 29th ACM International Conference on Multimedia*, 4034–4043.

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 286–301.

Supplementary Material:
Efficient Attention-Sharing Information Distillation Transformer
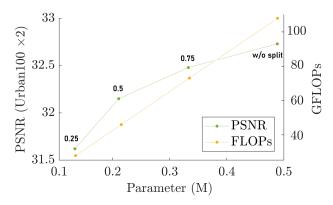# for Lightweight Single Image Super-Resolution

Figure 1: Ablation studies on channel-split methods. Results are evaluated on the Urban100 ($\times$2) dataset. The ratio represents how many channels from the entire feature are fed into self-attention layers.

Ablation Studies on Channel-Split

For ASID, we limit the number of feature channels fed into the spatial self-attention layer to reduce the model parameters and computational load. To evaluate the impact of channel-split on reconstruction performance, we analyze how changes in the number of channels fed into the self-attention layer influence overall performance. Fig. 1 illustrates the relationship between network performance and the ratio of channels processed by the self-attention module. Omni-SR (Wang et al. 2023) with three blocks is selected as our baseline for comparison. As demonstrated in Fig. 1, the computational load decreases proportionally with the ratio of channels. Meanwhile, the network's performance deteriorates at an increasing rate as a larger proportion of channels bypasses the self-attention layer. We set the channel-split ratio at 0.75 for the first IDB to balance the performance and network complexity.

Real Image Super-Resolution

Additionally, to evaluate the network's performance under unknown degradation rather than bicubic degradation, we perform training and evaluation on the RealSR (V3)

Table 1: Comparisons of SISR methods on RealSR dataset. **Bold** represents the proposed method.

Scale	Methods	PSNR	SSIM
$\times$3	IMDN	30.29	0.857
	LP-KPN	30.60	0.863
	ESRT	30.38	0.857
	ASID	**30.91**	**0.868**
$\times$4	IMDN	28.68	0.815
	LP-KPN	28.65	0.820
	ESRT	28.78	0.815
	ASID	**29.27**	**0.829**

dataset (Cai et al. 2019). During training, ASID models pre-trained on bicubic degradation are fine-tuned using the RealSR dataset. For the evaluation, we compare lightweight SR methods (Hui et al. 2019; Cai et al. 2019; Lu et al. 2022) applied to real-world super-resolution tasks. As shown in the table, despite not being specifically designed for the RealSR dataset, the proposed ASID consistently demonstrated strong performance across various metrics. These results highlight the potential of the proposed method to handle degradations beyond bicubic degradation and its effectiveness for real-world images.

Discussions

In this section, we discuss the advantages of the proposed ASID framework and its differences from previous methods.

ASID Framework

The benefits of the ASID framework can be summarized as follows. First, the information distillation scheme directly addresses the efficiency bottleneck of correlation matrix calculation, a common limitation in self-attention layers. The proposed structure reduces the number of channels involved in self-attention operations as the network deepens, alleviating computational inefficiencies. This enables ASID to stack multiple self-attention layers with fewer parameters and less computational cost compared to other Transformer-based SR methods. Second, attention-sharing significantly reduces the computational cost of self-attention layer operations. The spatial self-attention matrices, which define pixel relationships, are calculated in the 1st IDB and shared across

subsequent IDBs, eliminating a substantial portion of the self-attention operations. This strategy allows ASID to further reduce parameters and computational load while stacking self-attention layers. By combining these two strategies, ASID achieves high performance with substantially fewer operations and parameters than previous SR methods.

Comparison with Previous SR Methods

First, we review previous methods similar to ASID, both CNN-based and Transformer-based SR approaches, and highlight the key differences between them.

Information Distillation Methods The network design technique known as information distillation (Hui, Wang, and Gao 2018; Hui et al. 2019; Liu, Tang, and Wu 2020; Kong et al. 2022) is well-known for its efficient computation with minimal parameters, making it a popular choice in various CNN-based lightweight SR methods. Unlike traditional information distillation methods, which primarily enhance the efficiency of CNNs for handling local information, the Transformer-based ASID focuses on capturing long-range pixel dependencies through self-attention operations. Previous information distillation methods employ a series of calculation units that gradually reduce the channels involved in feature processing, thereby improving the efficiency of local feature extraction. In contrast, ASID integrates channel-split directly within the self-attention layers, rather than in separate modules dedicated to local feature extraction. This design maintains the Transformer's strength in capturing long-range dependencies. Moreover, we introduce an attention-sharing technique to address the significant computational bottleneck of self-attention layers in ASID. While previous information distillation methods use relatively lightweight calculation units due to their focus on local features, ASID faces an increased computational load as more self-attention layers are added. By sharing the self-attention matrix across layers, ASID reduces this computational burden without compromising network performance.

Transformer-based SR Methods Significant research has been conducted to enhance the practicality of Transformer-based SR methods. Zhang et al. (2022) proposed ELAN, which reduces network parameters and computational load by sharing the attention matrix. The primary difference between ELAN and ASID lies in their attention-sharing mechanisms and overall network design. ELAN shares a spatial attention matrix within the self-attention layers of the same building block. In contrast, our method computes all attention matrices in the first block, with subsequent blocks reusing these pre-computed matrices. This difference in sharing strategy enables ASID to allow adjacent layers to use different attention matrices, thereby capturing a more diverse set of pixel correlations compared to ELAN, where adjacent self-attention layers share the same spatial attention matrix. For a detailed comparison of these two attention-sharing strategies, please refer to the ablation study. Furthermore, the proposed information distillation structure in ASID enables the construction of a network with fewer parameters than ELAN, improving efficiency while

enhancing its ability to capture and utilize complex image details in SR tasks.

Next, Wang et al. (2023) proposed Omni Self-Attention (OSA), which considers pixel correlations at both meso-level and global-level by combining spatial and channel self-attention. While ASID shares a similar architecture with Omni-SR, a key difference lies in how attention is processed within the network. Omni-SR uses OSA to aggregate spatial and channel attention. In contrast, based on ablation studies from previous work (Wang et al. 2023), we determined that spatial self-attention contributes more significantly to Transformer performance than channel self-attention. Therefore, ASID separates spatial and channel attention, stacking multiple spatial self-attention layers. This architectural distinction means that Omni-SR calculates meso-level omni self-attention before global-level omni self-attention. In contrast, ASID computes spatial self-attention at both meso- and global-levels across the stacked spatial self-attention layers, with channel self-attention performed at the end of the module. Additionally, while Omni-SR relies solely on window-based self-attention to address the efficiency bottleneck, ASID achieves greater efficiency by incorporating the attention-sharing information distillation scheme. This not only mitigates computational challenges more effectively but also improves the overall performance of the lightweight Transformer-based SR network.

Qualitative Comparison

In Fig. 2, we compare zoomed-in results of previous lightweight SR methods and ASID on the Urban100 ($\times 4$) dataset. As shown in Fig. 2, ASID reconstructs SR images with precise detail while using the fewest model parameters, totaling only 313K. Notably, ASID produces fewer artifacts despite having significantly fewer parameters than other models. Furthermore, ASID restores SR images with quality comparable to Transformer-based methods with more than twice the parameters. This performance can be attributed to the proposed information distillation structure, which enables efficient stacking of self-attention layers and effectively expands the receptive field. This improvement is achieved without a substantial increase in parameters, underscoring the efficiency of ASID's lightweight design.

References

Cai, J.; Zeng, H.; Yong, H.; Cao, Z.; and Zhang, L. 2019. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3086–3095.

Hui, Z.; Gao, X.; Yang, Y.; and Wang, X. 2019. Lightweight image super-resolution with information multi-distillation network. In *Proceedings of the 27th ACM International Conference on Multimedia*, 2024–2032.

Hui, Z.; Wang, X.; and Gao, X. 2018. Fast and accurate single image super-resolution via information distillation network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 723–731.
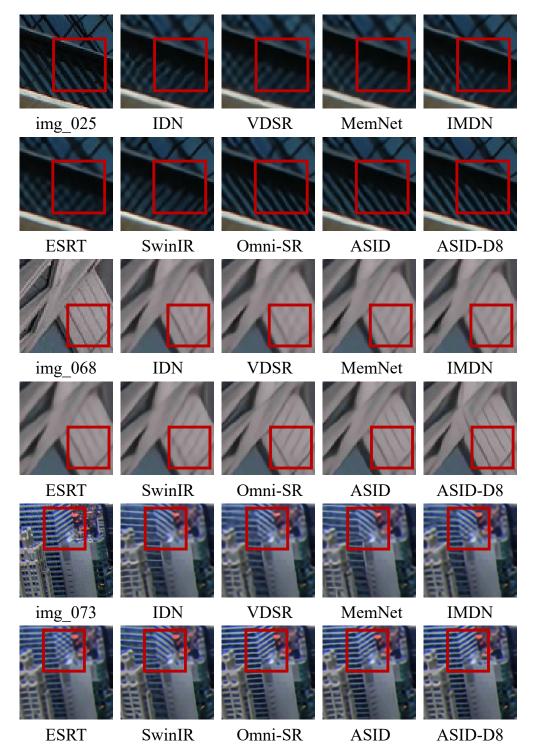
Figure 2: Qualitative Comparison of previous CNN-based and Transformer-based SR methods on Urban100 (×4) dataset. Note that ASID can restore accurate images with extremely few model parameters.

Kong, F.; Li, M.; Liu, S.; Liu, D.; He, J.; Bai, Y.; Chen, F.; and Fu, L. 2022. Residual local feature network for efficient super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 766–776.

Liu, J.; Tang, J.; and Wu, G. 2020. Residual feature distillation network for lightweight image super-resolution. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, 41–55. Springer.

Lu, Z.; Li, J.; Liu, H.; Huang, C.; Zhang, L.; and Zeng, T. 2022. Transformer for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 457–466.

Wang, H.; Chen, X.; Ni, B.; Liu, Y.; and Liu, J. 2023. Omni aggregation networks for lightweight image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22378–22387.

Zhang, X.; Zeng, H.; Guo, S.; and Zhang, L. 2022. Efficient long-range attention network for image super-resolution. In *European conference on computer vision*, 649–667. Springer.