DRCT: Saving Image Super-Resolution away from Information Bottleneck

Chih-Chung Hsu, Chia-Ming Lee, Yi-Shiuan Chou
Institute of Data Science, National Cheng Kung University
cchsu@gs.ncku.edu.tw, zuw408421476@gmail.com, nelly910421@gmail.com

Abstract

In recent years, Vision Transformer-based approaches for low-level vision tasks have achieved widespread success. Unlike CNN-based models, Transformers are more adept at capturing long-range dependencies, enabling the reconstruction of images utilizing non-local information. In the domain of super-resolution, Swin-transformer-based models have become mainstream due to their capability of global spatial information modeling and their shifting-window attention mechanism that facilitates the interchange of information between different windows. Many researchers have enhanced model performance by expanding the receptive fields or designing meticulous networks, yielding commendable results. However, we observed that it is a general phenomenon for the feature map intensity to be abruptly suppressed to small values towards the network's end. This implies an information bottleneck and a diminishment of spatial information, implicitly limiting the model's potential. To address this, we propose the Dense-residual-connected Transformer ***(DRCT)****, aimed at mitigating the loss of spatial information and stabilizing the information flow through dense-residual connections between layers, thereby unleashing the model's potential and saving the model away from information bottleneck. Experiment results indicate that our approach surpasses state-of-the-art methods on benchmark datasets and performs commendably at the NTIRE-2024 Image Super-Resolution (x4) Challenge. Our source code is available at https://github.com/ming053l/DRCT.*

1. Introduction

The task of Single Image Super-Resolution (SISR) is aimed at reconstructing a high-quality image from its low-resolution version. This quest for effective and skilled super-resolution algorithms has become a focal point of research within the field of computer vision, owing to its wide range of applications.

Following the foundational studies, CNN-based strategies [8, 19, 20, 35, 39, 40] have predominantly governed the super-resolution domain for an extended period. These strategies largely leverage techniques such as residual learning [12, 36, 46, 52, 61], or recursive learning [18, 21] for developing network architectures, significantly propelling the progress of super-resolution models forward.

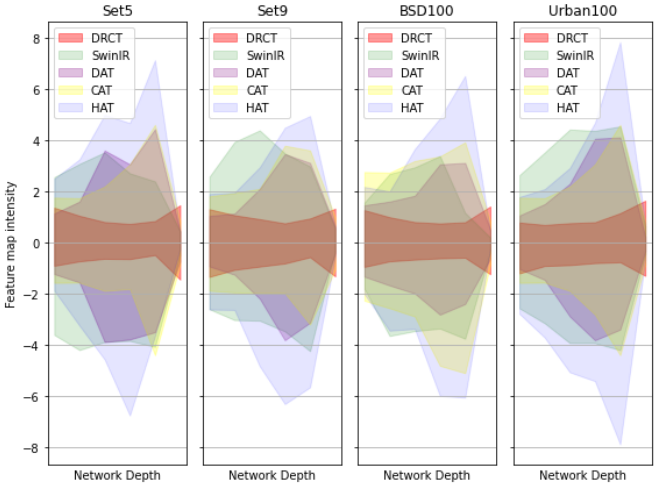

Figure 1. The feature map intensity on various benchmark datasets. We observed that feature map intensities decrease sharply at the end of SISR network, indicating potential information loss. In this paper, we propose DRCT to address this issue by enhancing receptive fields and adding dense-connections within residual blocks to mitigate information bottlenecks, thereby improving performance with a simpler model design.

CNN-based networks have achieved notable success in terms of performance. However, the inductive bias of CNN limits SISR models capture long-range dependencies. Their inherent limitations stem from the parameter-dependent scaling of the receptive field and the kernel size of convolution operator within different layers, which may neglect non-local spatial information within images.

To overcome the limitations associated with CNN-based networks, researchers have introduced Transformer-based SISR networks that leverage the capability to model long-range dependencies, thereby enhancing SISR performance. Notable examples include IPT [16] and EDT [33], which

utilize pre-training on large-scale dataset like ImageNet [14] to fully leverage the capabilities of Vision Transformer [9] for achieving ideal SISR results. Afterwards, SwinIR [34] incorporates Swin-Transformer [26] into SISR, marked a significant advancement in SISR performance.

This approach significantly enhances capabilities beyond those of traditional CNN-based models across various benchmarks. Following SwinIR's success, several works [4, 6, 32, 34, 58, 59, 63, 64] have built upon its framework. These subsequent studies leverage Transformers to innovate diverse network architectures specifically for super-resolution tasks, showcasing the evolving landscape of SISR technology through the exploration of new architectural innovations and techniques.

While using Transformer-based SISR model for inference across various datasets, we observed a common phenomenon: the intensity distribution of the feature maps undergoes more substantial changes as the network depth increases. This indicates the spatial information and attention intensity learned by the model. However, **there's often sharp decrease towards the end of the network** (refer to Figure 1) , shrinking to a smaller range. This phenomenon suggests that **such abrupt changes might be accompanied by a loss of spatial information**, indicating the presence of an information bottleneck.

Inspired by a series of works by Wang *et al.*, such as the YOLO-family [47, 50], CSPNet [48], and ELAN [49], we consider that network architectures based on SwinIR, despite significantly enlarging the receptive fields through shift-window attention mechanism to address the small receptive fields in CNNs, are prone to gradient bottlenecks due to the loss of spatial information as network depth increases. This implicitly constrains the model's performance and potential.

To address the issue of spatial information loss due to an increased number of network layers, we introduce the Dense-residual-connected Transformer (DRCT), designed to stabilize the forward-propagation process and prevent information bottlenecks. This is achieved by the proposed Swin-Dense-Residual-Connected Block (SDRCB), which incorporates Swin Transformer Layers and transition layers into each Residual Deep feature extraction Group (RDG). Consequently, this approach enhances the receptive field with fewer parameters and a simplified model architecture, thereby resulting in improved performance. The main contributions of this paper are summarised as follows:

- We observed that as the network depth increases, the intensity of the feature map will gradually increase, then abruptly drop to a smaller range. This severe oscillation may be accompanied by information loss.
- We propose DRCT by adding dense-connection within residual groups to stabilize the information flow for deep feature extraction during propagation, thereby saving the SISR model away from the information bottleneck.
- By integrating dense connections into the Swin-Transformer-based SISR model, the proposed DRCT achieves state-of-the-art performance while maintaining efficiency. This approach showcases its potential and raises the upper-bound of the SISR task.

2. Related works

2.1. Vision Transformer-based Super-Resolution

IPT [16], a versatile model utilizing the Transformer encoder-decoder architecture, has shown efficacy in several low-level vision tasks. SwinIR [34], building on the Swin Transformer [26] encoder, employs self-attention within local windows during feature extraction for larger receptive fields and greater performance, compared to traditional CNN-based approaches. UFormer [53] introduces an innovative local-enhancement window Transformer block, utilizing a learnable multi-scale restoration modulator within the decoder to enhance the model's ability to detect both local and global patterns. ART [64] incorporates an attention retractable module to expand its receptive field, thereby enhancing SISR performance. CAT [5] leverages rectangle-window self-attention for feature aggregation, achieving a broader receptive field. HAT [4] integrates channel attention mechanism [51] with overlapping cross-attention module, activating more pixels to reconstruct better SISR results, thereby setting new benchmarks in the field.

2.2. Auxiliary Supervision and Feature Fusion

Auxiliary Supervision. Deep supervision is a commonly used auxiliary supervision method [13, 31] that involves training by adding prediction layers at the intermediate levels of the model [47–49]. This approach is particularly prevalent in architectures based on Transformers that incorporate multi-layer decoders. Another popular auxiliary supervision technique involves guiding the feature maps produced by the intermediate layers with relevant metadata to ensure they possess attributes beneficial to the target task [11, 28, 30, 52, 61]. Choosing the appropriate auxiliary supervision mechanism can accelerate the model's convergence speed, while also enhancing its efficiency and performance.

Feature Fusion. Many studies have explored the integration of features across varying dimensions or multi-level features, such as FPN [22], to obtain richer representations for different tasks [36, 55]. In CNNs, attention mechanisms have been applied to both spatial and channel dimensions to improve feature representation; examples of which include RTCS [10] and SwinFusion [37]. In ViT [9], spatial self-attention is used to model the long-range dependencies between pixels. Additionally, some researchers have investigated the incorporation of channel attention within Trans-

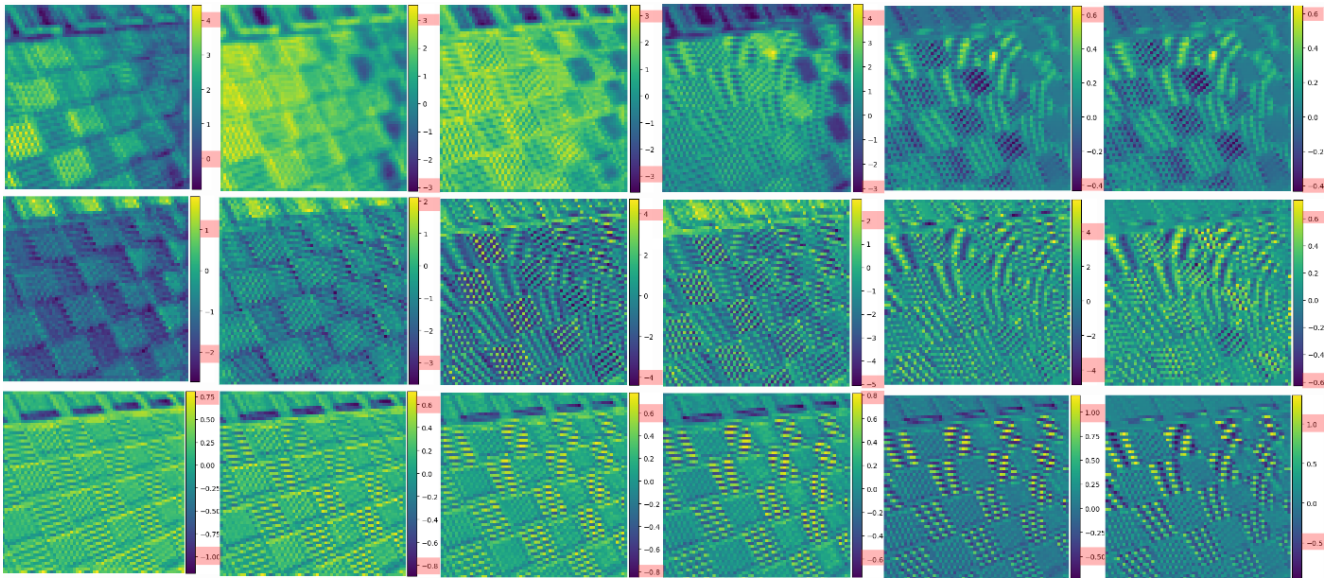

Figure 2. The feature map visualization displays, from top to bottom, SwinIR [34], HAT [4], and the proposed DRCT, with positions further to the right representing deeper layers within the network. For both SwinIR and HAT, the intensity of the feature maps is significant in the shallower layers but diminishes towards the network's end. We consider this phenomenon implies the loss of spatial information, leading to the limitation and information bottleneck with SISR tasks. As for the proposed DRCT, the learned feature maps are gradually and stably enhanced without obvious oscillations. It represents the stability of the information flow during forward propagation, thereby yielding higher intensity in the final layer's output. (zoom in to better observe the color-bar besides feature maps.)

formers [3, 62] to effectively amalgamate spatial and channel information, thereby improving model performance.

3. Problem Statement

3.1. Information Bottleneck Principle

According to the information bottleneck principle [45], the given data X may cause information loss when going through consecutive layers. It may lead to gradient vanish when back-propagation for fitting network parameters and predicting Y, as shown in the equation below:

$$I(X, X) \geq I(Y, X) \geq I(Y, f_\theta(X)) \geq I(X, g_\phi(f_\theta(X))), \tag{1}$$

where I indicates mutual information, f and g are transformation functions, and θ and ϕ are parameters of f and g, respectively.

In deep neural networks, $f_\theta(\cdot)$ and $g_\phi(\cdot)$ respectively represent the two consecutive layers in neural network. From equation (1), we consider that as the number of network layer becomes deeper, the information flow will be more likely to be lost. In term of SISR tasks, the general goal is to find the mapping function F with optimized function parameters θ to maximize the mutual information between HR and SR image.

$$F(\mathbf{I}_{LR}; \theta) = \mathbf{I}_{SR}; \max_\theta I(\mathbf{I}_{HR}; F(\mathbf{I}_{LR}; \theta)) \tag{2}$$

3.2. Spatial Information Vanish in Super-resolution

Generally speaking, SISR methods [4–6, 32, 34, 58, 63, 64] can generally divided into three parts: (1) shallow feature extraction, (2) deep feature extraction, (3) image reconstruction. Among these methods, there is almost no difference between shallow feature extraction and image reconstruction. The former is composed of simple convolution layers, while the latter consists of convolution layers and upsampling layers. However, deep feature extraction differs significantly. Yet, their commonality lies in being composed of various residual blocks, which can be simply defined as:

$$X^{l+1} = X^l + f_\theta^{l+1}(X^l), \tag{3}$$

where X indicates inputs, f is a consecutive layers for l'th residual group , and θ represents the parameters of f^l.

Especially for SISR task, two methods of stabilizing information flow or training process are introduced:

Residual connection to learn local feature. Adopting residual learning allows the model to only update the differences between layers, rather than output the total information from a previous layer directly [12]. This reduces the difficulty of model training and prevents gradient vanishing locally [61]. However, according to our observations, while this design effectively transmits spatial information between different residual blocks, there may still be infor-

mation loss.

Because the information within a residual block may not necessarily maintain spatial information, this ultimately leads to non-smoothness in terms of feature map intensity (refer to Fig. 2), causing an information bottleneck at the deepest layers during forward propagation. This makes it easy for spatial information to be lost as the gradient flow reaches the deeper layers of the network, resulting in reduced data efficiency or the need for more complex network designs to achieve better performance.

Dense connection to stabilize information flow. Incorporating dense connections into the Swin-Transformer based SISR model offers two significant advantages. Firstly, *global auxiliary supervision*. It effectively fuses the spatial information across different residual groups [52, 61], preserving high-frequency features throughout the deep feature extraction. Secondly, *saving SISR model away from information bottleneck*. By leveraging the integration of spatial information, the model ensures a smooth transmission of spatial information [46], thereby mitigating the information loss and enhancing the receptive field.

4. Motivation

Dense-Residual Group auxiliary supervision. Motivated by RRDB-Net [52], Wang *et al.* suggested that incorporating dense-residual connections can aggregate multi-level spatial information and stabilize the training process [35, 41]. We consider that it is possible to stabilize the information flow within each residual-groups during propagation, thereby saving SISR model away from the information bottleneck.

Dense connection with Shifting-window mechanism. Recent studies on SwinIR-based methods have concentrated on enlarging the receptive field [4–6, 64] by sophisticated WSA or enhancing the network's capability to extract features [32, 53] for high-quality SR images. By adding dense-connections [15] within Swin-Transformer-based blocks [26, 34] in the SISR network for deep feature extraction, the proposed DRCT's receptive field is enhanced while capturing long-range dependencies. Consequently, this approach allows for achieving outstanding performance with simpler model architectures [46], or even using shallower SISR networks.

5. Methodology

5.1. Network Architecture

As shown in Figure 3, DRCT comprises three distinct components: shallow feature extraction, deep feature extraction, and image reconstruction module, respectively.

Shallow and deep feature extraction. Given a low-resolution (LR) input $\mathbf{I}_{LR} \in \mathbb{R}^{H\times W\times C_{in}}$ (H, W and C_{in} are the image height, width and input channel number, respectively), we use a 3×3 convolution layer Conv$(\cdot)$ [54] to extract shallow feature $\mathbf{F}_0 \in \mathbb{R}^{H\times W\times C}$ as

$$\mathbf{F}_0 = \text{Conv}(\mathbf{I}_{LQ}), \tag{4}$$

Then, we extract deep feature which contains high-frequency spatial information $\mathbf{F}_{DF} \in \mathbb{R}^{H\times W\times C}$ from $\mathbf{F}_0$ and it can be defined as

$$\mathbf{F}_{DF} = H_{DF}(\mathbf{F}_0), \tag{5}$$

where $H_{DF}(\cdot)$ is the deep feature extraction module and it contains K Residual Deep feature extraction Group (RDG) and single convolution layer Conv$(\cdot)$ for feature transition. More specifically, intermediate features $\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_K$ and the output deep feature $\mathbf{F}_{DF}$ are extracted block by block as

$$\mathbf{F}_i = \text{RDG}_i(\mathbf{F}_{i-1}), \quad i = 1, 2, \ldots, K, \tag{6}$$

$$\mathbf{F}_{DF} = \text{Conv}(\mathbf{F}_K), \tag{7}$$

Image reconstruction. We reconstruct the SR image $\mathbf{I}_{SR} \in \mathbb{R}^{H\times W\times C_{in}}$ by aggregating shallow and deep features, it can be defined as:

$$\mathbf{I}_{SR} = H_{\text{rec}}(\mathbf{F}_0 + \mathbf{F}_{DF}), \tag{8}$$

where $H_{\text{rec}}(\cdot)$ is the function of the reconstruction for fusing high-frequency deep feature $\mathbf{F}_{DF}$ and low-frequency feature $\mathbf{F}_0$ together to obtain SR result.

5.2. Deep Feature Extraction

Residual Deep feature extraction Group. In developing of RDG, we take cues from RRDB-Net [52] and RDN [61], employing a residual-dense block (RDB) as the foundational unit for SISR. The reuse of feature maps emerges as the enhanced receptive field in the RDG's feed-forward mechanism. To expound further, RDG with several SDRCB enhances the capability to integrate information across different scales, thus allowing for a more comprehensive feature extraction. RDG facilitates the information flow within residual group, capturing the local features and spatial information group by group.

Swin-Dense-Residual-Connected Block. In purpose of capturing the long-range dependency, we utilize the shifting window self-attention mechanism of Swin-Transformer Layer (STL) [26, 34] for obatining adaptive receptive fields, complementing RRDB-Net by focusing on multi-level spatial information. This synergy leverages STL to dynamically adjust the focus of the model based on the global content of the input, allowing for a more targeted and efficient extraction of features. This mechanism ensures that even as the depth of the network increases, global details

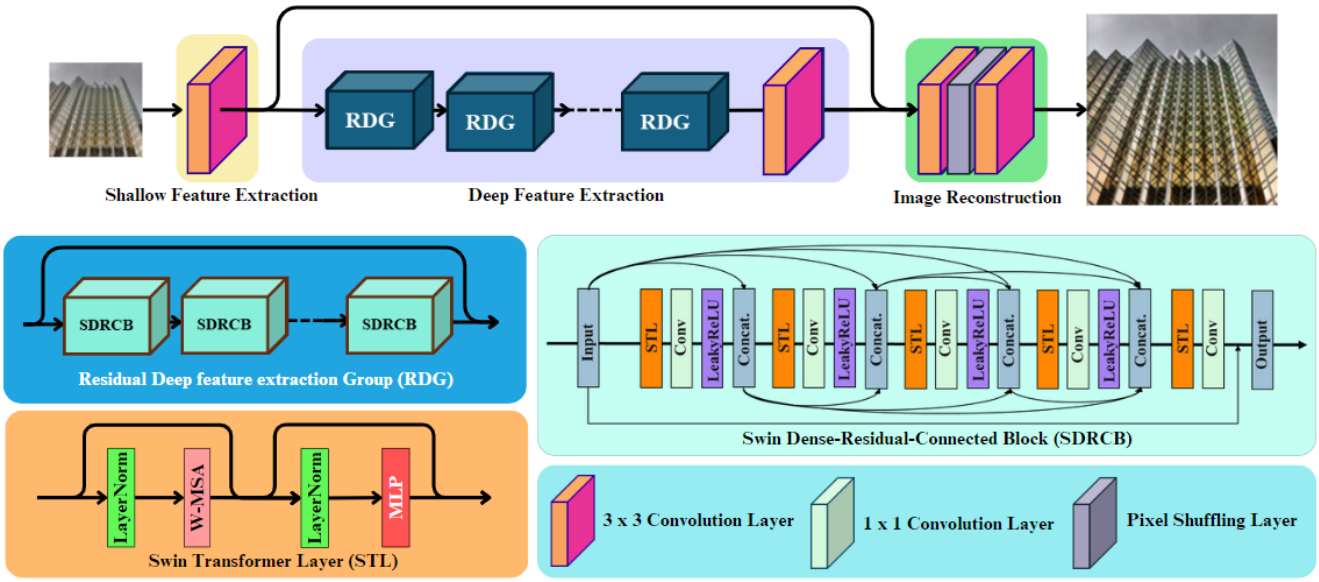

Figure 3. The overall architecture of the proposed DRCT and the structure of RDG and SDRCB.

are preserved, effectively enlarging and enhancing the receptive field without compromising. By integrating STLs with dense-residual connections, the architecture benefits from both a vast receptive field and the capability to hone in on the most relevant information, thereby enhancing the model's performance in SISR tasks requiring detailed and context-aware processing. For the input feature maps $\mathbf{Z}$ within RDG, the SDRCB can be defined as:

$$\mathbf{Z}_j = H_{\text{trans}}(\text{STL}([\mathbf{Z}, ..., \mathbf{Z}_{j-1}])), j = 1, 2, 3, 4, 5, \quad (9)$$

$$\text{SDRCB}(\mathbf{Z}) = \alpha \cdot \mathbf{Z}_5 + \mathbf{Z}, \quad (10)$$

where $[\cdot]$ denotes the concatenation of multi-level feature maps produced by the previous layers. $H_{\text{trans}}(\cdot)$ refers to the convolution layer with a LeakyReLU activate function for feature transition. The negative slope of LeakyReLU is set to 0.2. Conv_1 is the 1×1 convolution layer, which is used to adaptively fuse a range of features with different levels [42]. α represents residual scaling factor, which is set to 0.2 for stabilizing the training process [52].

5.3. Same-task Progressive Training Strategy

In recent years, Progressive Training Strategy (PTS) [17, 25] has gained increased attention and can be seen as a method of fine-tuning. Compared to conventional training methods, PTS tends to converge model parameters to more desirable local minima. HAT [4] introduces the Same-task Pre-training, which aims to train the model on a large dataset like ImageNet [14] before fine-tuning it on a specific dataset, leading to improved SISR results. Lei *et al.* [57] proposed initially training a SISR network with L1-loss and then using L2-loss to eliminate artifacts, achieving better results on the PSNR metric. This has been widely adopted [64]. We proposed a Same-task Progressive Training Strategy (SPTS). At first, we pre-trained DRCT on ImageNet to initialize model parameters and then fine-tuned on specific datasets with L1 loss,

$$\ell_{L1} = \|I_{HR} - I_{SR}\|_1, \quad (11)$$

and finally use L2 loss to eliminate singular pixels and artifacts, therefore further archiving greater performance on PSNR metric.

$$\ell_{L2} = \|I_{HR} - I_{SR}\|_2 \quad (12)$$

6. Experiment Results

6.1. Dataset

Our DRCT model is trained on DF2K, a substantial aggregated dataset that includes DIV2K [1] and Flickr2K [44]. DIV2K provides 800 images for training, while Flickr2K contributes 2650 images. For the training input, we generate LR versions of these images by applying a bicubic downsampling method with scaling factors of 2, 3, and 4, respectively. To assess the effectiveness of our model, we conduct performance evaluations using well-known SISR benchmark datasets such as Set5 [2], Set14 [56], BSD100 [38], Urban100 [29], and Manga109 [24].

Method	Scale	Training Dataset	Set5 [2]		Set14 [56]		BSD100 [38]		Urban100 [29]		Manga109 [24]	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR [35]	×2	DIV2K	38.11	0.9602	33.92	0.9195	32.32	0.9013	32.93	0.9351	39.10	0.9773
RCAN [60]	×2	DIV2K	38.27	0.9614	34.12	0.9216	32.41	0.9027	33.34	0.9384	39.44	0.9786
SAN [20]	×2	DIV2K	38.31	0.9620	34.07	0.9213	32.42	0.9028	33.10	0.9370	39.32	0.9792
IGNN [19]	×2	DIV2K	38.24	0.9613	34.07	0.9217	32.41	0.9025	33.23	0.9383	39.35	0.9786
HAN [40]	×2	DIV2K	38.27	0.9614	34.16	0.9217	32.41	0.9027	33.35	0.9385	39.46	0.9785
NLSN [39]	×2	DIV2K	38.34	0.9618	34.08	0.9231	32.43	0.9027	33.42	0.9394	39.59	0.9789
SwinIR [34]	×2	DF2K	38.42	0.9623	34.46	0.9250	32.53	0.9041	33.81	0.9427	39.92	0.9797
CAT-A [5]	×2	DF2K	38.51	0.9626	34.78	0.9265	32.59	0.9047	34.26	0.9440	40.10	0.9805
HAT [4]	×2	DF2K	38.63	0.9630	34.86	0.9274	32.62	0.9053	34.45	0.9466	40.26	0.9809
DAT [6]	×2	DF2K	38.58	0.9629	34.81	0.9272	32.61	0.9051	34.37	0.9458	40.33	0.9807
DRCT (Ours)	×2	DF2K	**38.72**	**0.9646**	**34.96**	**0.9287**	**32.75**	**0.9071**	**34.54**	**0.9474**	**40.41**	**0.9814**
IPT† [16]	×2	ImageNet	38.37	-	34.43	-	32.48	-	33.76	-	-	-
EDT† [33]	×2	DF2K	38.63	0.9632	34.80	0.9273	32.62	0.9052	34.27	0.9456	40.37	0.9811
HAT-L† [4]	×2	DF2K	38.91	0.9646	35.29	0.9293	32.74	0.9066	35.09	0.9505	41.01	0.9831
DRCT-L‡ (Ours)	×2	DF2K	**39.14**	**0.9658**	**35.36**	**0.9302**	**32.90**	**0.9078**	**35.17**	**0.9516**	**41.14**	**0.9842**
EDSR [35]	×3	DIV2K	34.65	0.9280	30.52	0.8462	29.25	0.8093	28.80	0.8653	34.17	0.9476
RCAN [60]	×3	DIV2K	34.74	0.9299	30.65	0.8482	29.32	0.8111	29.09	0.8702	34.44	0.9499
SAN [20]	×3	DIV2K	34.75	0.9300	30.59	0.8476	29.33	0.8112	28.93	0.8671	34.30	0.9494
IGNN [19]	×3	DIV2K	34.72	0.9298	30.66	0.8484	29.31	0.8105	29.03	0.8696	34.39	0.9496
HAN [40]	×3	DIV2K	34.75	0.9299	30.67	0.8483	29.32	0.8110	29.10	0.8705	34.48	0.9500
NLSN [39]	×3	DIV2K	34.85	0.9306	30.70	0.8485	29.34	0.8117	29.25	0.8726	34.57	0.9508
SwinIR [34]	×3	DF2K	34.97	0.9318	30.93	0.8534	29.46	0.8145	29.75	0.8826	35.12	0.9537
CAT-A [5]	×3	DF2K	35.06	0.9326	31.04	0.8538	29.52	0.8160	30.12	0.8862	35.38	0.9546
HAT [4]	×3	DF2K	35.07	0.9329	31.08	0.8555	29.54	0.8167	30.23	0.8896	35.53	0.9552
DAT [6]	×3	DF2K	35.16	0.9331	31.11	0.8550	29.55	0.8169	30.18	0.8886	35.59	0.9554
DRCT (Ours)	×3	DF2K	**35.18**	**0.9338**	**31.24**	**0.8569**	**29.68**	**0.8182**	**30.34**	**0.8910**	**35.76**	**0.9575**
IPT† [16]	×3	ImageNet	34.87	-	30.85	-	29.38	-	29.49	-	-	-
EDT† [33]	×3	DF2K	35.13	0.9328	31.09	0.8553	29.53	0.8165	30.07	0.8863	35.47	0.9550
HAT-L† [4]	×3	DF2K	35.28	0.9345	31.47	0.8584	29.63	0.8191	30.92	0.8981	36.02	0.9576
DRCT-L‡ (Ours)	×3	DF2K	**35.32**	**0.9348**	**31.54**	**0.8591**	**29.68**	**0.8211**	**31.14**	**0.9004**	**36.16**	**0.9585**
EDSR [35]	×4	DIV2K	32.46	0.8968	28.80	0.7876	27.71	0.7420	26.64	0.8033	31.02	0.9148
RCAN [60]	×4	DIV2K	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087	31.22	0.9173
SAN [20]	×4	DIV2K	32.64	0.9003	28.92	0.7888	27.78	0.7436	26.79	0.8068	31.18	0.9169
IGNN [19]	×4	DIV2K	32.57	0.8998	28.85	0.7891	27.77	0.7434	26.84	0.8090	31.28	0.9182
HAN [40]	×4	DIV2K	32.64	0.9002	28.90	0.7890	27.80	0.7442	26.85	0.8094	31.42	0.9177
NLSN [39]	×4	DIV2K	32.59	0.9000	28.87	0.7891	27.78	0.7444	26.96	0.8109	31.27	0.9184
SwinIR [34]	×4	DF2K	32.92	0.9044	29.09	0.7950	27.92	0.7489	27.45	0.8254	32.03	0.9260
CAT-A [5]	×4	DF2K	33.08	0.9052	29.18	0.7960	27.99	0.7510	27.89	0.8339	32.39	0.9285
HAT [4]	×4	DF2K	33.04	0.9056	29.23	0.7973	28.00	0.7517	27.97	0.8368	32.48	0.9292
DAT [6]	×4	DF2K	33.08	0.9055	29.23	0.7973	28.00	0.7515	27.87	0.8343	32.51	0.9291
DRCT (Ours)	×4	DF2K	**33.11**	**0.9064**	**29.35**	**0.7984**	**28.18**	**0.7532**	**28.06**	**0.8378**	**32.59**	**0.9304**
IPT† [16]	×4	ImageNet	32.64	-	29.01	-	27.82	-	27.26	-	-	-
EDT† [33]	×4	DF2K	32.82	0.9031	29.09	0.7939	27.91	0.7483	27.46	0.8246	32.05	0.9254
HAT-L† [4]	×4	DF2K	33.30	0.9083	29.47	0.8015	28.09	0.7551	28.60	0.8498	33.09	0.9335
DRCT-L‡ (Ours)	×4	DF2K	**33.37**	**0.9090**	**29.54**	**0.8025**	**28.16**	**0.7577**	**28.70**	**0.8508**	**33.14**	**0.9347**

Table 1. Quantitative comparison with the several peer-methods on benchmark datasets. "†" indicates that methods adopt pre-training strategy [4] on ImageNet. "‡" represents that methods use same-task progressive-training strategy. The top three results are marked in **red**, **blue**, and **orange** , respectively.

6.2. Implementation Details

The training process can be structured into three phases, as Section 4-3 illustrates. (1) pre-trained on ImageNet [14], (2) optimize the model on the given dataset, (3) L2-loss for PSNR enhancement. Throughout the training process, we use the Adam optimizer with $\beta_1 = 0.9$, and $\beta_2 = 0.999$ and train for $800k$ iterations in the first and second stages.

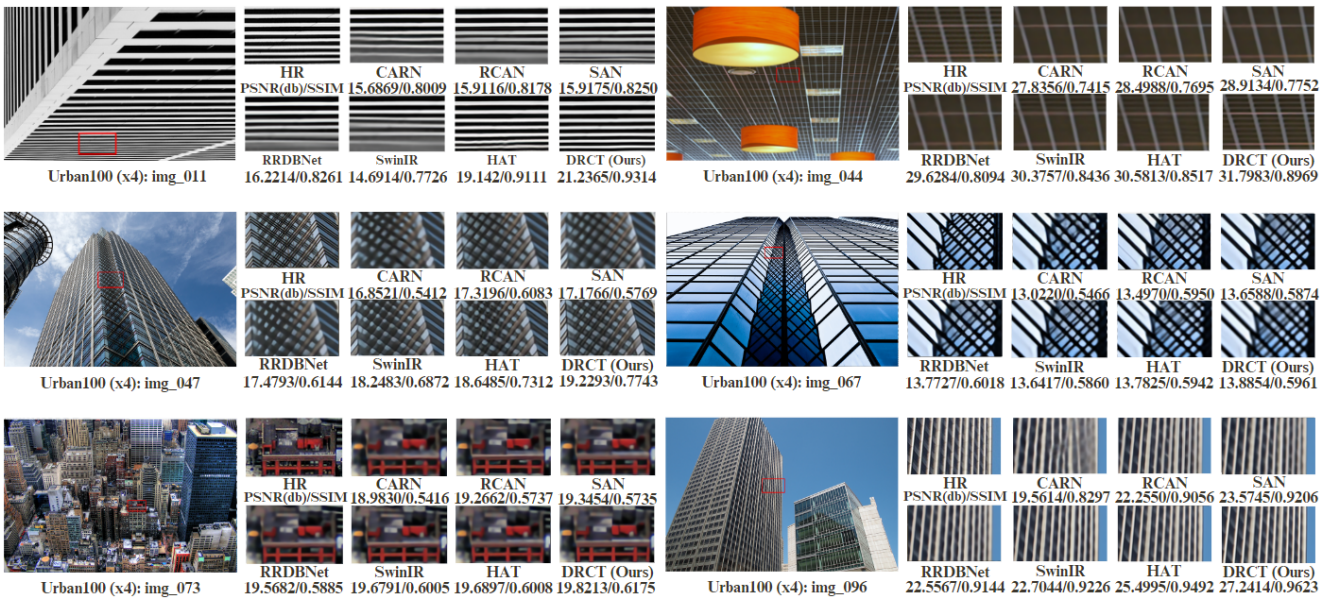

Figure 4. Visual comparison on × 4 SISR. The patches for comparison are marked with red boxes in the original images. The higher the PSNR/SSIM metrics, the better the performance..

The learning rate is set to $2e-4$, and the multi-step learning scheduler is also used. The learning rate is halved at the $300k$, $500k$, $650k$, $700k$, $750k$ iterations respectively. Weight decay is not applied, and the batch size is set to 32. In the architecture of DRCT, the configuration of depth and width is maintained identically to that of HAT [4]. To elaborate, both the number of RDG and SDRCB units are established at 6, and the channel number of intermediate feature maps is designated as 180. The attention head number and window size are set to 6 and 16 for window-based multi-head self-attention (W-MSA). In terms of data preparation, HR patches with dimensions of 256×256 pixels were extracted from the HR images. To improve the generalizability, we apply random horizontal flips and rotation augmentation.

6.3. Quantitative Results

For the evaluation, we use full RGB channels and ignore the $(2 \times \text{scale})$ pixels from the border. PSNR and SSIM metrics are used to evaluation criteria. Table 1 presents the quantitative comparison of our approach and the state-of-the-art methods, including EDSR [35], RCAN [60], SAN [20], IGN [19], HAN [40], NLSN [39], SwinIR [34], CAT-A [5], DAT [6], as well as approaches using ImageNet pretraining, such as IPT [16], EDT [33] and HAT [4]. We can see that our method outperforms the other methods significantly on all benchmark datasets. In addition, the DRCT-L can bring further improvement and greatly expand the performance upper-bound on SISR tasks. Even with fewer model parameters and computational requirements, DRCT is also significantly greater than the state-of-the-art methods.

6.4. Visual Comparison

The visual comparisons displayed in Figure 4. For the selected images from Urban100 [29], DRCT is effective in restoring structures, whereas other methods suffer from notably blurry effects. The visual results demonstrate the superiority of our approach.

Along with providing visualizations for the LAM [27], we compute the Diffusion Index (DI), which is the attribution-based analysis. The DI reflects the range of involved pixels. A higher DI refers to a wider range of attention. In scenarios where DRCT used fewer parameters (which will be discussed in the next subsection), it achieves a higher DI. This outcome suggests that, after enhancing the receptive field through SDRCB, the model can leverage a long-range dependency and non-local information for SISR without the need for intricately designed W-MSA.

6.5. Model Complexity

To demonstrate the potential of our proposed DRCT, we conducted further analysis on model complexity and performance.

Model efficiency. In Table 2, the proposed DRCT clearly requires fewer computational resources compared to HAT in terms of parameter size, multiply-add operations, memory requirements, and FLOPs. Specifically, when scaling up the model sizes of DRCT and HAT, DRCT-L surpasses HAT-L in all metrics.

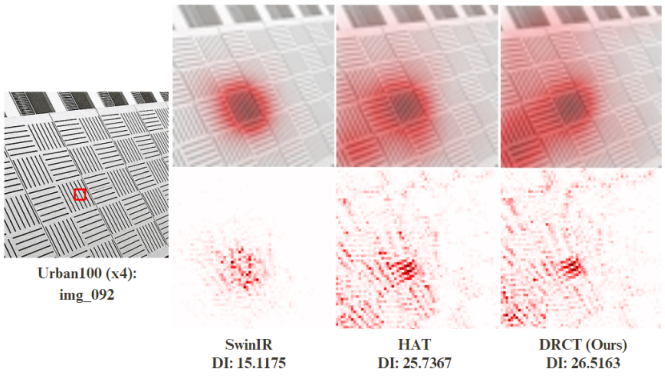

Figure 5. The LAM [27] visualization. DRCT improves performance by enhancing the receptive field to mitigate the issue of spatial information loss in deeper layers of the network.

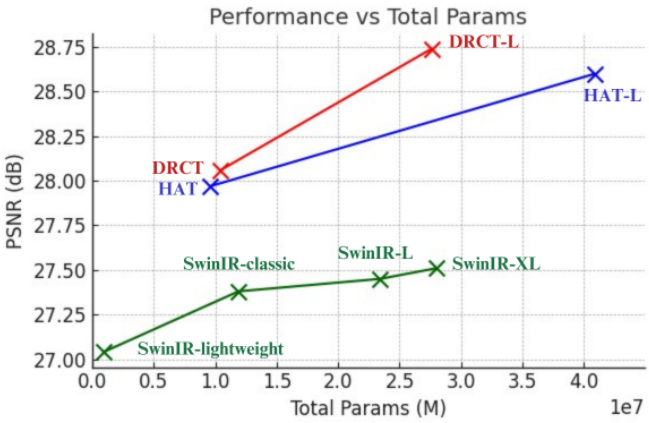

Figure 6. The model complexity comparison between SwinIR, HAT, and proposed DRCT evaluated on Urban100 [29] dataset.

Model performance. From Figure 6, we can observe that the performance curves of the HAT and SwinIR models are approaching horizontal lines, suggesting that the performance is nearing a bottleneck and its upper-bound, even if scaling up the model parameters.

This demonstrates that the design of DRCT, which incorporates dense-connections in the residual groups within a Swin-transformer-based model to stabilize the information flow, achieves convincing results with a reduced computational burden.

	#Params.	#Multi-Adds.	Forward or Backward pass	FLOPs
HAT [4]	**9.621M**	11.22G	2053.42M	42.18G
DRCT	10.443M	**5.92G**	**1857.55M**	**7.92G**
HAT-L [4]	40.846M	76.69G	5165.39M	79.60G
DRCT-L	**27.580M**	**9.20G**	**4278.19M**	**11.07G**

Table 2. Model complexity analysis for (×4) SISR on Urban100.

6.6. NTIRE Image Super-Resolution (x4) Challenge

	Validation phase	Testing phase
PSNR	31.1820	31.1776
SSIM	0.8494	0.8620

Table 3. NTIRE 2024 Challenge Results with x4 SR in terms of PSNR and SSIM on validation phase and testing phase.

The dataset for the NTIRE 2024 Image Super-Resolution (x4) Challenge [7] comprises three collections: DIV2K [1], Flickr2K [44], and LSDIR [23]. Specifically, the DIV2K dataset provides 800 pairs of HR and LR images for training. The LR images are obtained from the HR images after bicubic downsampling with specific scaling factor. For validation, it offers 100 LR images for the purpose of creating SR images, with the HR versions to be made available at the challenge's final stage. Additionally, the test dataset includes 100 varied LR images. The self-ensemble strategy is used for testing-time augmentation (TTA) [43]. Our TTA methods include random rotation, and horizontal and vertical flipping. We also conducted a model ensemble strategy for fusing different reconstructed results by HAT [4] and the proposed DRCT to eliminate the annoying artifacts and improve final SR quality. Our SISR model was entered into both the validation and testing phases of this challenge, with the detailed in Table 3.

7. Conclusion

In this paper, we introduce the phenomenon of information bottlenecks observed in SISR models, where spatial information is lost as network depth increases during forward propagation. This may lead to information loss when limiting the upper bound of model performance for the SISR task, which requires detailed spatial information and context-aware processing.

To address these issues, we present a novel Swin-transformer-based model, Dense-residual-connected Transformer (DRCT). The design philosophy behind DRCT centers on stabilizing the information flow and enhancing the receptive fields by incorporating dense-connections within residual blocks, combining the shift-window attention mechanism to adaptively capture global information.

As a result, the model can better focus on global spatial information and surpass existing state-of-the-art methods without the need for designing sophisticated window attention mechanisms or increasing model parameters. The experiment results have demonstrated the efficacy of the proposed DRCT, indicating its effectiveness and the potential for future work related to SISR tasks.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 5, 8

[2] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012. 5, 6

[3] Qiang Chen, Qiman Wu, Jian Wang, Qinghao Hu, Tao Hu, Errui Ding, Jian Cheng, and Jingdong Wang. Mixformer: Mixing features across windows and dimensions, 2022. 3

[4] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer, 2023. 2, 3, 4, 5, 6, 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, 6, 7

[6] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xiaokang Yang, and Fisher Yu. Dual aggregation transformer for image super-resolution. In *ICCV*, 2023. 2, 3, 4, 6, 7

[7] Zheng Chen, Zongwei Wu, Eduard-Sebastian Zamfir, Kai Zhang, Yulun Zhang, Radu Timofte, Xiaokang Yang, et al. Ntire 2024 challenge on image super-resolution (x4): Methods and results. In *Computer Vision and Pattern Recognition Workshops*, 2024. 8

[8] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks, 2015. 1

[9] Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 2

[10] Chih-Chung Hsu et al. Real-time compressed sensing for joint hyperspectral image transmission and restoration for cubesat. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 2

[11] Chaoxu Guo et al. Augfpn: Improving multi-scale feature learning for object detection, 2019. 2

[12] Christian Ledig et al. Photo-realistic single image super-resolution using a generative adversarial network, 2017. 1, 3

[13] Christian Szegedy et al. Going deeper with convolutions, 2014. 2

[14] Deng Jia et al. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2, 5, 6

[15] Gao Huang et al. Densely connected convolutional networks, 2018. 4

[16] Hanting Chen et al. Pre-trained image processing transformer, 2021. 1, 2, 6, 7

[17] Keyao Wang et al. Dynamic feature queue for surveillance face anti-spoofing via progressive training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 5

[18] Ryan Dahl et al. Pixel recursive super resolution, 2017. 1

[19] Shangchen Zhou et al. Cross-scale internal graph neural network for image super-resolution. In *Advances in Neural Information Processing Systems*, 2020. 1, 6, 7

[20] Tao Dai et al. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11065–11074, 2019. 1, 6, 7

[21] Tai Ying et al. Image super-resolution via deep recursive residual network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1

[22] Tsung-Yi Lin et al. Feature pyramid networks for object detection, 2017. 2

[23] Yawei Li et al. Lsdir: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1775–1787, 2023. 8

[24] Yusuke Matsui et al. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76 (20):21811–21838, 2017. 5, 6

[25] Yulun Zhang et al. Ntire 2023 challenge on image super-resolution (×4): Methods and results. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1865–1884, 2023. 5

[26] Ze Liu et al. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. 2, 4

[27] Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9199–9208, 2021. 7, 8

[28] Zeeshan Hayder, Xuming He, and Mathieu Salzmann. Boundary-aware instance segmentation, 2017. 2

[29] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5197–5206, 2015. 5, 6, 7, 8

[30] Kuan-Chih Huang, Tsung-Han Wu, Hung-Ting Su, and Winston H. Hsu. Monodtr: Monocular 3d object detection with depth-aware transformer, 2022. 2

[31] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-Supervised Nets. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 562–570, San Diego, California, USA, 2015. PMLR. 2

[32] Ao Li, Le Zhang, Yun Liu, and Ce Zhu. Feature modulation transformer: Cross-refinement of global representation via high-frequency prior for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12514–12524, 2023. 2, 3, 4

[33] Wenbo Li, Xin Lu, Shengju Qian, Jiangbo Lu, Xiangyu Zhang, and Jiaya Jia. On efficient transformer and image pre-training for low-level vision. *arXiv preprint arXiv:2112.10175*, 2021. 1, 6, 7

[34] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. *arXiv preprint arXiv:2108.10257*, 2021. 2, 3, 4, 6, 7

[35] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 1, 4, 6, 7

[36] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1, 2

[37] Jiayi Ma, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, and Yong Ma. Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7):1200–1217, 2022. 2

[38] 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*, pages 416–423 vol.2, 2001. 5, 6

[39] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3517–3526, 2021. 1, 6, 7

[40] 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, 2020. 1, 6, 7

[41] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning, 2016. 4

[42] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *Proceedings of International Conference on Computer Vision*, 2017. 5

[43] Radu Timofte, Rasmus Rothe, and Luc Van Gool. Seven ways to improve example-based single image super resolution, 2015. 8

[44] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, Lei Zhang, Bee Lim, et al. Ntire 2017 challenge on single image super-resolution: Methods and results. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 5, 8

[45] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5, 2015. 3

[46] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image super-resolution using dense skip connections. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 4

[47] Chien-Yao Wang and Hong-Yuan Mark Liao. YOLOv9: Learning what you want to learn using programmable gradient information. 2024. 2

[48] Chien-Yao Wang, Hong-Yuan Mark Liao, I-Hau Yeh, Yueh-Hua Wu, Ping-Yang Chen, and Jun-Wei Hsieh. Cspnet: A new backbone that can enhance learning capability of cnn, 2019. 2

[49] Chien-Yao Wang, Hong-Yuan Mark Liao, and I-Hau Yeh. Designing network design strategies through gradient path analysis. *arXiv preprint arXiv:2211.04800*, 2022. 2

[50] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7464–7475, 2023. 2

[51] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks, 2020. 2

[52] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *The European Conference on Computer Vision Workshops (ECCVW)*, 2018. 1, 2, 4, 5

[53] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration, 2021. 2, 4

[54] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better, 2021. 4

[55] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022. 2

[56] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces*, pages 711–730, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. 5, 6

[57] Lei Zha, Yu Yang, Zicheng Lai, Ziwei Zhang, and Juan Wen. A lightweight dense connected approach with attention on single image super-resolution. *Electronics*, 10:1234, 2021. 5

[58] Dafeng Zhang, Feiyu Huang, Shizhuo Liu, Xiaobing Wang, and Zhezhu Jin. Swinfir: Revisiting the swinir with fast fourier convolution and improved training for image super-resolution, 2023. 2, 3

[59] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution, 2022. 2

[60] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 6, 7

[61] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2472–2481, 2018. 1, 2, 3, 4

[62] Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Anima Anandkumar, Jiashi Feng, and Jose M. Alvarez. Understanding the robustness in vision transformers, 2022. 3

[63] Yupeng Zhou, Zhen Li, Chun-Le Guo, Song Bai, Ming-Ming Cheng, and Qibin Hou. Srformer: Permuted self-attention for single image super-resolution. *arXiv preprint arXiv:2303.09735*, 2023. 2, 3

[64] Qiang Zhu, Pengfei Li, and Qianhui Li. Attention retractable frequency fusion transformer for image super resolution. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1756–1763, 2023. 2, 3, 4, 5