GYF

This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.

Except for this watermark, it is identical to the accepted version;

the final published version of the proceedings is available on IEEE Xplore.

**Feature Modulation Transformer: Cross-Reﬁnement of Global Representation**

**via High-Frequency Prior for Image Super-Resolution**

Ao Li1 , Le Zhang1 \*, Yun Liu2 , Ce Zhu1

1University of Electronic Science and Technology of China, 2I2R, A\*STAR

aoli@std .uestc .edu .cn, {lezhang,eczhu}@uestc .edu .cn, vagrantlyun@gmail .com

**Abstract**

*Transformer-based methods have exhibited remarkable potential in single image super-resolution (SISR) by effec- tively extracting long-range dependencies. However, most of the current research in this area has prioritized the de- sign of transformer blocks to capture global information, while overlooking the importance of incorporating high- frequency priors, which we believe could be beneﬁcial. In our study, we conducted a series of experiments and found that transformer structures are more adept at cap- turing low-frequency information, but have limited capacity in constructing high-frequency representations when com- pared to their convolutional counterparts. Our proposed solution, the* ***c****ross-****r****eﬁnement* ***a****daptive* ***f****eature modulation* ***t****ransformer (****CRAFT****), integrates the strengths of both con- volutional and transformer structures. It comprises three key components: the* ***high-frequency enhancement resid- ual block (HFERB)****for extracting high-frequency informa- tion, the* ***shift rectangle window attention block (SRWAB)*** *for capturing global information, and the* ***hybrid fusion block (HFB)*** *for reﬁning the global representation. Our experiments on multiple datasets demonstrate that CRAFT outperforms state-of-the-art methods by up to* ***0.29dB*** *while using fewer parameters. The source code will be made available at:* [https://github. com/AVC2-UESTC/](https://github.com/AVC2-UESTC/) CRAFT-SR .git*.*

**1. Introduction**

Single image super-resolution (SISR) has garnered sig- niﬁcant attention in recent years, owing to its promising applications across diverse domains, such as surveillance video and medical image enhancement [31, 10], old im- age reconstruction [21, 17], and efﬁcient image transmis- sion [47]. Despite its practical value, SISR remains an ill- posed problem, given the existence of multiple solutions for a given low-resolution (LR) image. To tackle this challenge,

\*Corresponding author.

a multitude of classical approaches have been proposed, in- cluding A+ [36], SC [41], and ANR [35]. However, these methods exhibit limitations in their performance, primarily attributed to their constrained model capacities.

In recent years, deep learning has experienced signiﬁcant growth and demonstrated remarkable success in SISR [7, 20, 45, 22]. Prior research efforts have introduced resid- ual and dense connectives to facilitate the stacking of deep convolutional neural networks (CNNs) [16, 37], while oth- ers [46, 40, 29, 30] have leveraged attention mechanisms to enhance performance. Notably, the emergence of trans- former architectures has demonstrated their efﬁcacy in cap- turing long-range dependencies and attaining state-of-the- art performance [21, 6, 4, 18, 25]. Despite these advance- ments, these works have mainly focused on designing trans- former blocks to obtain global information and overlooked the potential of incorporating high-frequency priors [32, 8] to further bolster performance in SISR. Additionally, there is limited detailed analysis of the impact of frequency on performance.

In this paper, we investigate the inﬂuence of high- frequency information on the performance of CNN and transformer structures in SISR. We achieve this by discard- ing different ratios of high-frequency components from the input image and observing the corresponding performance changes. Our empirical ﬁndings reveal that transformers tend to prioritize low-frequency information and exhibit limited capability in constructing high-frequency represen- tations when compared to CNNs. To address this issue, we proposed a cross-reﬁnement adaptive feature modulation transformer (CRAFT) that integrates the strengths of both structures. Speciﬁcally, CRAFT comprises three key com- ponents, namely the high-frequency enhancement residual block (HFERB), the shift rectangle window attention block (SRWAB), and the hybrid fusion block (HFB), which work collaboratively to capture high-frequency details, extract long-range dependencies, and reﬁne the output for better representation. Experimental results show that CRAFT out- performs state-of-the-art performance with relatively fewer parameters. The main contributions of this paper are as fol-

lows:

• We study the impact of CNN and transformer struc- tures on performance from a frequency perspective and observe that transformer is more effective in capturing low-frequency information while having limited ca- pacity for constructing high-frequency representations compared to CNN.

• Based on the observation, we design a parallel struc- ture to explore different frequency features. We utilize the HFERB branch to introduce high-frequency infor- mation, which is beneﬁcial to SISR, and the SRWAB branch to acquire global information.

• We propose a fuse strategy that integrates the strengths of CNN and transformer. Speciﬁcally, we treat the HFERB branch as high-frequency prior and the output of SRWAB as key and value for inter-attention, result- ing in improved performance.

• Extensive experimental results on multiple datasets show that the proposed method performs on par with the existing state-of-the-art SISR methods while using fewer parameters.

**2. Related Works**

**2.1. CNN-based SISR**

Since the pioneering work SRCNN [7] has achieved sig- niﬁcant progress in SISR, various CNN-based works have been proposed. Kim *et al*. [15] presented an SR method using deep networks by cascading 20 layers, demonstrating promising results. Building upon this, Lim *et al*. [23] in- troduced the enhanced deep super-resolution (EDSR) net- work, which achieved a signiﬁcant performance boost by removing the batch normalization layer [14] from the resid- ual block and incorporating additional convolution layers. Ahn *et al*. [2] designed an architecture with an increased number of residual blocks and dense connections, further improving the SR performance. In pursuit of lightweight models, Hui *et al*. [13] proposed a selective fusion ap- proach, employing cascaded information multi-distillation blocks to construct an efﬁcient model. Li *et al*. [19] intro- duced a method involving predeﬁned ﬁlters and utilized a CNN to learn coefﬁcients, which were then linearly com- bined to obtain the ﬁnal results. Sun *et al*. [34] proposed a hybrid pixel-unshufﬂed network (HPUN) by introducing an efﬁcient and effective downsampling module into the SR task.

**2.2. Transformer-based SISR**

Liang *et al*. [21] proposed SwinIR, a robust baseline model for image restoration, leveraging the Swin Trans-

former [24]. CAT [6] modiﬁed the window shape and in- troduced a rectangle window attention to obtaining better performance. Chen *et al*. [4] proposed a pre-trained image processing transformer and showed that pre-trained mecha- nism could signiﬁcantly improve the performance for low- level tasks. Li *et al*. [18] comprehensively analyzed the effect of pre-training and proposed a versatile model to tackle different low-level tasks. Lu *et al*. [25] proposed a lightweight transformer to capture long-range dependencies between similar patches in an image with the help of the specially designed efﬁcient transformer and efﬁcient atten- tion mechanism. Zhang *et al*. [44] introduced a shift con- volution and a group-wise multi-scale self-attention to re- duce the complexity of transformer. HAT [5] introduced a hybrid attention mechanism to enhance the performance of window-based transformers.

**3. Analysis of Frequency Impact**

This section delves into the inﬂuence of performance from a frequency perspective. To analyze the impact of var- ious frequencies on CNN and transformer, we conduct two sets of experiments using four common used benchmarks, as illustrated in Figure 1.

We select CARN [2], IMDN [13], EDSR [23], and SwinIR [21], CAT [6], HAT [5] as representatives of CNN and transformer structures. The process of dropping fre- quency components is depicted in Figure 1(c). Given a high-resolution (HR) image X HR , we perform a fast Fourier transform (FFT) on it to obtain its frequency spec- trum. Subsequently, we ﬂatten this spectrum into a se- quence and arrange it in ascending order based on the mag- nitudes. With a sequence length of L, we deﬁne a threshold determined by the drop ratio γ , 0 ≤ γ ≤ 1, located at the magnitude corresponding to the position γ · L. Frequency components with magnitudes below this threshold are set to zero. Following this, we perform an inverse fast Fourier transform (IFFT) to generate the HR image with dropped

frequencies, referred to as X(γ). The formulation for

this process is as follows

X(γ) = IFFT (Drop(|FFT (XHR )|,γ)). (1)

Afterward, we downsample X(γ) using bicubic in-

terpolation to obtain the LR version Xrp (γ) (*e.g*.

×4 down-sampling). Finally, we employ CNN-based and transformer-based SR models to generate the super-

resolved counterpart Xrp (γ).

To analyze the dependency of CNN and transformer on high-frequency information, we compute the peak

signal-to-noise ratio (PSNR) PD (γ) between Xrp (γ) and

X . We then plot the PSNR drop trend to visualize the

difference between the two structures. As shown in Fig- ure 1(a), the PSNR drop ratio for each drop ratio is deﬁned

-3

-4

-5

-6

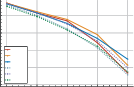
-7

Set5

-2

-4

-6

HAT

CAT

SwinIR CARN

-8

IMDN

EDSR

-10

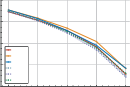
70 75 80 85 90

Drop Ratio (%)

Set14

-4

-6

HAT

-8

-10

CAT

SwinIR CARN IMDN EDSR

-12

70 75 80 85 90

Drop Ratio (%)

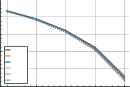
BSD100

-2

-3

-4

-5

HAT

CAT

-6

-7

SwinIR CARN

IMDN

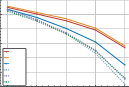
EDSR

-8

70 75 80 85 90

Drop Ratio (%)

Urban100

HAT

CAT

SwinIR CARN IMDN EDSR

70 75 80 85 90

Drop Ratio (%)

(a) Dependency of different structures on high-frequency information.

0

-0.5

PSNR Drop Ratio (%)

-1

-1.5

-2

-2.5

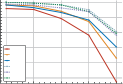
Set5

0

PSNR Drop Ratio (%)

-0.5

-1

HAT

CAT

SwinIR CARN IMDN EDSR

-1.5

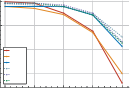
70 75 80 85 90

Drop Ratio (%)

Set14

PSNR Drop Ratio (%)

-0.5

HAT

-1

CAT

SwinIR CARN

IMDN

EDSR

-1.5

70 75 80 85 90

Drop Ratio (%)

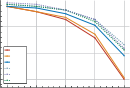
BSD100

0

PSNR Drop Ratio (%)

-1

-2

HAT

CAT

SwinIR CARN IMDN EDSR

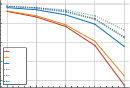
-3

-4

70 75 80 85 90

Drop Ratio (%)

Urban100

HAT

CAT

SwinIR CARN IMDN EDSR

70 75 80 85 90

Drop Ratio (%)

(b) Effectiveness of reconstructing high-frequency information.

PSNR Drop Ratio (%)

Ratio (%)

PSNR Drop

PSNR Drop Ratio (%)

PSNR Drop Ratio (%)





Dropping

Sampling

SR

IFFT

FFT

HR Spectrum of HR Dropped Spectrum Dropped HR Dropped LR Dropped SR

of HR

(c) The procedure of dropping high-frequency.

Figure 1. The inﬂuence of high-frequency information on the performance of CNN and transformer architectures. Dashed and solid lines correspond to CNN and transformer methods, respectively. (a) With an increase in the high-frequency drop ratio, transformer models exhibit a smaller change in PSNR compared to CNN, suggesting their superiority in capturing low-frequency information. (b) As the high-frequency drop ratio increases, transformer models show a more pronounced change in PSNR compared to CNN, indicating their limited ability to reconstruct high-frequency information from low-frequency.

as

Rop (γ) =  , (2)

where P (0) represents the PSNR without dropping, cal- culated between XSR and X HR . The ﬁgures illustrate that the transformer model exhibits reduced sensitivity to high-frequency information and excels in capturing low- frequency information, as evidenced by the smaller PSNR change compared to the CNN model as the proportion of discarded high-frequency information increases.

Furthermore, we conduct another experiment to evalu- ate the effectiveness of different structures in reconstruct- ing high-frequency information. Speciﬁcally, we calculate

the PSNR PE (γ) between Xrp (γ) and XHR and plot the

performance drop trend as previously depicted. The PSNR drop ratio for each drop ratio can be expressed as

Rop (γ) =  . (3)

From Figure 1(b), we observe that as the proportion of discarded high-frequency information increases, the trans- former model experiences a larger PSNR change compared

to the CNN model, indicating its limited ability to recon- struct high-frequency information from low-frequency.

Based on these observations, we argue that the trans- former requires the assistance of CNN to enhance its ca- pability to recover intricate details. To address this, we pro- pose a method that combines the strengths of both CNN and transformer. Speciﬁcally, we introduce CNN information as a high-frequency prior to aid the transformer in reﬁning the global representation.

**4. Proposed Method**

The CRAFT network comprises three key components: Shallow feature extraction, residual cross-reﬁnement fusion groups (RCRFGs), and reconstruction as shown in Figure 2. The shallow feature extraction module comprises a single convolutional layer, while the reconstruction module is fol- lowed by the SwinIR [21]. TheRCRFG component consists of several cross-reﬁnement fusion blocks (CRFBs), each comprising three types of blocks: the high-frequency en- hancement residual blocks (HFERBs), the shift rectangle window attention blocks (SRWABs), and the hybrid fusion blocks (HFBs). We ﬁrst describe the overall structure of



**Reconstruction**

|  |
| --- |
|  |

**RCRFG**

|  |
| --- |
|  |

**RCRFG**

|  |
| --- |
|  |

onv

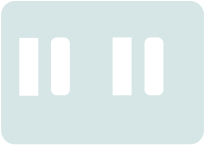
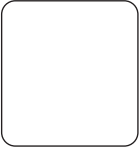
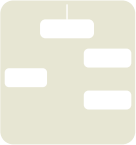
|  |
| --- |
|  |

onv

|  |
| --- |
|  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Residual Cross-refinement Fusion Group**  **(RCRFG)** | | | |
| |  | | --- | | **CRFB** | | |  | | --- | | **CRFB** | | |  | | --- | | onv | |  |

|  |
| --- |
| **SRWAB** |
|  |



Rwin-SA

|  |
| --- |
|  |

LayerNorm

|  |
| --- |
|  |

LayerNorm

|  |
| --- |
|  |

MLP

**CRFB**

XS



C

C

C

**HFERB**

**SRWAB**

**SRWAB**

**HFB**

C

XH

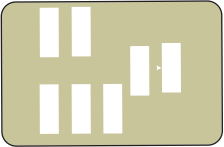
 Element-wise Multiplication  Element-wise Addition

 GELU Activation

 Fixed-function

 Trainable

|  |
| --- |
|  |



|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |

C

C/2



**HFERB**

C/2

|  |
| --- |
| **Hybrid Fusion Block (HFB)** |
| aerNorv  onv onv 3×3  DWConv  3×3 3×3  DWConv DWConv  Inter-Attention  LayerNorm  onv onv  3×3 3×3  DWConv DWConv  onv   |  | | --- | |  |  |  | | --- | |  |  |  | | --- | |  |  |  | | --- | |  |   XH   |  | | --- | |  |  |  | | --- | |  |  |  | | --- | |  |  |  | | --- | |  |  |  | | --- | |  |   XS   |  | | --- | |  |  |  | | --- | |  | |

Figure 2. The framework of CRAFT. HFERB extracts the high frequency from the input features, SRWAB captures the long-range depen- dency of input features, and HFB integrates the output of HFERB and SRWAB to cross reﬁne the global features. Best viewed in color.

CRAFT and then elaborate on the three key designs, includ- ing HFERB, SRWAB, and HFB.

**4.1. Model Overview**

The input LR image is processed by a 3 × 3 convolutional layer to obtain shallow features. These features are then fed into a serial of RCRFGs to learn deep features. After the last RCRFG, a 3 × 3 convolutional layer aggregates the features, and a residual connection is established between its output and the shallow features for facilitating training. The reconstruction module employs a 3 × 3 convolutional layer to aggregate the features, and a shufﬂe layer [33] is used to obtain the ﬁnal SR output image.

**4.2. High-frequency Enhancement Residual Block**

The HFERB aims to enhance the high-frequency infor- mation, as shown in Figure 2. It comprises the local feature extraction (LFE) branch and the high-frequency enhance- ment (HFE) branch. Speciﬁcally, we split the input features

Fin ∈RH ×W ×C into two parts, and then processed by the two branches separately

FiFE , FiFE = Split(Fin ), (4)

where FiFE , FiFE ∈ RH ×W ×C/2 represent the input of

LFE and HFE. For the LFE branch, we utilize a 3 × 3 con- volutional layer followed by a GELU activation function to extract local high-frequency features

iFE = fa (Conv3 ×3 (FiFE )), (5)

where the Conv3 ×3 (·) refers to the convolutional layer and the fa (·) represents the GELU activation layer. For the HFE branch, we employ a max-pooling layer to extract

high-frequency information from the input features FiFE .

Then, we use a 1 × 1 convolutional layer followed by a GELU activation function to enhance the high-frequency features,

iFE = fa (Conv1 ×1 (MaxPooling(FiFE ))), (6)

where the Conv1 ×1 (·) indicates the convolutional layer, the

MaxPooling(·) means the max-pooling layer and the fa (·) represents the GELU activation layer. The outputs of the two branches are then concatenated and fed into a 1 × 1 convolutional layer to fuse the information thoroughly. To make the network beneﬁt from multi-scale information and maintain training stability, a skip connection is introduced. The whole process can be formulated as

XH = Conv1 ×1 (Concat(iFE , iFE )) + Fin , (7)

where the Concat(·) refers to the concatenation operation and the Conv1 ×1 (·) represents the convolutional layer.

**4.3. Shift Rectangle Window Attention Block**

We utilize the shift rectangle window (SRWin) to ex- pand the receptive ﬁeld, which can beneﬁt SISR [6]. Un- like square windows, the SRWin uses rectangle windows to capture more relevant information along the longer axis. In detail, given an input Xin ∈ RH×W ×C, we divide it into



height and width of the rectangle window. For the i-th rect- angle window feature Xi ∈ R(rh ×rw)×C, we compute the *query*, *key*, and *value* as follows

**Q**i = Xi WiQ , **K**i = Xi WiK , **V**i = Xi WiV , (8)

where the WiQ∈RC×d , WiK ∈RC×d and WiV ∈RC×d repre- sent the projection matrices and d is projection dimension

which is commonly set to d =  M is the num- ber of heads. The self-attention can be formulated as

 (9)

where B is the dynamic relative position encoding [38]. Moreover, a convolutional operation on the *value* is intro- duced to enhance local extraction capability. To capture in- formation from different axes, we utilize two types of rect- angle windows: Horizontal and vertical windows. Unlike traditional operations that utilize attention masks to limit calculations to the same window, in practice, we eliminate the mask and enable more extensive information interac- tion across different windows. Accordingly, we split the attention heads into two equal groups and compute the self- attention within each group separately. We then concatenate the outputs of the two groups to obtain the ﬁnal output. The procedure can be expressed as

Rwin-SA(X) = Concat(V-Rwin, H-Rwin)Wp , (10)

where the Wp ∈ RC×C represents the linear projection to fuse the features, V-Rwin and H-Rwin indicate the verti- cal and horizontal rectangle window attention. In addition,

a multi-layer perceptron (MLP) is used for further feature transformations. The whole process can be formulated as

X = Rwin-SA(LN (Xin )) + Xin

(11)

XS = MLP (LN (X)) + X,

where the LN represents the LayerNorm layer.

**4.4. Hybrid Fusion Block**

To better integrate the merits of CNN and transformer (HFERB and SRWAB), we have designed a hybrid fusion block (HFB), which is illustrated in Figure 2. We for- mulate the output of HFERB as the high frequency prior *query* and the output of SRWAB as *key*, *value* and calcu- late the inter-attention to reﬁne the global features which are obtained from SRWAB. Moreover, most existing meth- ods focus on spatial relations and overlook channel infor- mation. To overcome this limitation, we perform inter- attention based on the channel dimension to explore channel dependencies. This design will signiﬁcantly reduce com- plexity. Traditional methods that utilize spatial attention tend to result in signiﬁcant computational complexity (*e.g*., O(N2 C), N ≥ C), where N represents the length of the sequence and C represents the channel dimension. In con- trast, our channel attention design can transfer the quadratic component to the channel dimension (*e.g*., O(NC2 )), ef- fectively reducing complexity.

Speciﬁcally, as shown in Figure 2, we use a 1 × 1 convo- lutional layer followed by a 3 × 3 depth-wise convolutional layer to generate the high frequency query **Q** ∈ RH×W ×C based on the output of HFERB, XH . As to the output of SRWAB, XS , we ﬁrst normalize the features by LayerNorm layer and then use the same operation as the query **Q** to get the key **K** ∈RH×W ×C and the value **V** ∈RH×W ×C . Fol- lowing the [42], we perform the reshape operation on **Q**,

**K** and **V** to get the ∈ RC×(HW ) , ∈ RC×(HW ) and

∈RC×(HW ) . After that, we compute the inter-attention

as

 (12) where the α represents the learnable parameter. Meanwhile, we add the reﬁnement features to the XS to get the fu- sion output Xfuse. In addition, we feed Xfuse to an im- proved feed-forward network [42] to aggregate the features further. The details of this structure are shown in Figure 2. It introduced a gate mechanism to fully extract the spatial and channel information and gain better performance. The whole process can be formulated as

Xfuse =Inter-Atten(LN (XS ), XH )+XS

(13)

XHFB =IM LP (LN (X))+Xfuse ,

where the LN means LayerNorm operation, IM LP rep- resents the improved MLP, and Inter-Atten indicates

Table 1. Performance comparison of different SISR models on ﬁve benchmarks. Params represents the total number of network parameters.

Results for the best and second best candidates are **highlighted**, and underlined.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Scale | Model | Params | Set5  (PSNR/SSIM) | Set14  (PSNR/SSIM) | BSD100  (PSNR/SSIM) | Urban100  (PSNR/SSIM) | Manga109  (PSNR/SSIM) |
| ×2 | EDSR-baseline [23]  CARN [2]  IMDN [13]  LatticeNet [26]  LAPAR-A [19]  HPUN-L [34] | 1370K  1592K  694K  756K  548k  714K | 37.99/0.9604  37.76/0.9590  38.00/0.9605  38.06/0.9607  38.01/0.9605  38.09/0.9608 | 33.57/0.9175  33.52/0.9166  33.63/0.9177  33.70/0.9187  33.62/0.9183  33.79/0.9198 | 32.16/0.8994  32.09/0.8978  32.19/0.8996  32.20/0.8999  32.19/0.8999  32.25/0.9006 | 31.98/0.9272  31.92/0.9256  32.17/0.9283  32.25/0.9288  32.10/0.9283  32.37/0.9307 | 38.54/0.9769  38.36/0.9765  38.88/0.9774  -/-  38.67/0.9772  39.07/0.9779 |
| SwinIR-light [21]  ESRT [25]  ELAN-light [44]  CRAFT (Ours) | 878K  777K  582K  737K | 38.14/0.9611  38.03/0.9600  38.17/0.9611  **38.23**/**0.9615** | 33.86/0.9206  33.75/0.9184  **33.94**/0.9207  33.92/**0.9211** | 32.31/0.9012  32.25/0.9001  32.30/0.9012  **32.33**/**0.9016** | 32.76/0.9340  32.58/0.9318  32.76/0.9340  **32.86**/**0.9343** | 39.12/0.9783  39.12/0.9774  39.11/0.9782  **39.39**/**0.9786** |
| ×3 | EDSR-baseline [23]  CARN [2]  IMDN [13]  LatticeNet [26]  LAPAR-A [19]  HPUN-L [34] | 1555K  1592K  703K  765K  544k  723K | 34.37/0.9270  34.29/0.9255  34.36/0.9270  34.53/0.9281  34.36/0.9267  34.56/0.9281 | 30.28/0.8417  30.29/0.8407  30.32/0.8417  30.39/0.8424  30.34/0.8421  30.45/0.8445 | 29.09/0.8052  29.06/0.8034  29.09/0.8046  29.15/0.8059  29.11/0.8054  29.18/0.8072 | 28.15/0.8527  28.06/0.8493  28.17/0.8519  28.33/0.8538  28.15/0.8523  28.37/0.8572 | 33.45/0.9439  33.50/0.9440  33.61/0.9445  -/-  33.51/0.9441  33.90/0.9463 |
| SwinIR-light [21]  LBNet [9]  ESRT [25]  ELAN-light [44]  CRAFT (Ours) | 886K  736K  770K  590K  744K | 34.62/0.9289  34.47/0.277  34.42/0.9268  34.61/0.9288  **34.71**/**0.9295** | 30.54/0.8463  30.38/0.8417  30.43/0.8433  30.55/0.8463  **30.61**/**0.8469** | 29.20/0.8082  29.13/0.8061  29.15/0.8063  29.21/0.8081  **29.24**/**0.8093** | 28.66/0.8624  28.42/0.8559  28.46/0.8574  28.69/0.8624  **28.77**/**0.8635** | 33.98/0.9478  33.82/0.9406  33.95/0.9455  34.00/0.9478  **34.29**/**0.9491** |
| ×4 | EDSR-baseline [23]  CARN [2]  IMDN [13]  LatticeNet [26]  LAPAR-A [19]  HPUN-L [34] | 1518K  1592K  715K  777K  659k  734K | 32.09/0.8938  32.13/0.8937  32.21/0.8948  32.18/0.8943  32.15/0.8944  32.31/0.8962 | 28.58/0.7813  28.60/0.7806  28.58/0.7811  28.61/0.7812  28.61/0.7818  28.73/0.7842 | 27.57/0.7357  27.58/0.7349  27.56/0.7353  27.57/0.7355  27.61/0.7366  27.66/0.7386 | 26.04/0.7849  26.07/0.7837  26.04/0.7838  26.14/0.7844  26.14/0.7871  26.27/0.7918 | 30.35/0.9067  30.47/0.9084  30.45/0.9075  -/-  30.42/0.9074  30.77/0.9109 |
| SwinIR-light [21]  LBNet [9]  ESRT [25]  ELAN-light [44]  CRAFT (Ours) | 897K  742K  751K  601K  753K | 32.44/0.8976  32.29/0.8960  32.19/0.8947  32.43/0.8975  **32.52**/**0.8989** | 28.77/0.7858  28.68/0.7832  28.69/0.7833  28.78/0.7858  **28.85**/**0.7872** | 27.69/0.7406  27.62/0.7382  27.69/0.7379  27.69/0.7406  **27.72**/**0.7418** | 26.47/0.7980  26.27/0.7906  26.39/0.7962  26.54/0.7982  **26.56**/**0.7995** | 30.92/0.9151  30.76/0.9111  30.75/0.9100  30.92/0.9150  **31.18**/**0.9168** |

the proposed inter-attention mechanism, which introduces high-frequency prior to reﬁning the global representations.

**5. Experiments**

**5.1. Data and Metrics**

In this paper, we adopt the DIV2K [1] as the training dataset, which includes 800 training images. Meanwhile, ﬁve benchmarks are used for evaluation, including Set5 [3], Set14 [43], BSD100 [27], Urban100 [12], and Manga109 [28] with three magniﬁcation factors: ×2, ×3, and ×4. The quality of the images is evaluated using PSNR, and SSIM [39]. The complexity of the model is indicated by its pa- rameters.

**5.2. Implementation Details**

Following the general setting, we use bicubic to obtain the corresponding LR images from the original HR images. During training, we randomly crop the images into 64 × 64 patches, and the total training iterations are 500K. Mean- while, data augmentation is performed, such as random hor- izontal ﬂipping and 90。rotation. The Adam optimizer with β1 = 0.9 and β2 = 0.999 is adopted to minimize the L1 Loss. The batch size is set to 64, the initial learning

rate is set to 2 × 10−4 and reduced by half at the mile- stone [250K, 400K, 450K, 475K]. In addition, the model is trained on 4 NVIDIA 3090 GPUs using the PyTorch tool- box. In CRAFT, we have set the RCRFG number to 4 and the CRFB number to 2 for each RCRFG. Each CRFB is comprised of 1 HFERB and 2 SRWABs for efﬁciency. The feature channel, attention head, and MLP expansion ratio are set to 48, 6, and 2, respectively. We also set the IMLP expansion ratio to 2.66, as in [42]. To obtain two types of rectangle windows, we have set the rectangle window size to [sh, sw] as [4, 16] and [16, 4].

**5.3. Comparison with state-of-the-art methods**

We compare with several state-of-the-art SISR meth- ods to demonstrate the effective of the proposed CRAFT model, including EDSR [23], CARN [2], IMDN [13], LatticeNet [26], LAPAR [19], SwinIR [21], HPUN [34], ESRT [25], LBNet [9], and ELAN [44].

**Quantitative Results.** The experimental results for SISR are presented in Table 1, where the proposed CRAFT model demonstrates competitive performance across all benchmarks. Particularly, when compared to traditional CNN-based methods like EDSR, the proposed CRAFT achieves signiﬁcant performance improvements of 0.85dB,



Img012 ( ×4)



Bicubic



SwinIR [21]



EDSR [23]



ESRT [25]



CARN [2]



CRAFT (Ours)



LBNet [9]



HR





Bicubic





EDSR [23]





CARN [2]





LBNet [9]



YumeiroCook ( ×4) SwinIR [21] ESRT [25] CRAFT (Ours) HR





Bicubic





EDSR [23]





CARN [2]





LBNet [9]



UchuKigekiM774 ( ×4) SwinIR [21] ESRT [25] CRAFT (Ours) HR

Figure 3. Visual quality comparison with SOTA methods. CRAFT achieves better restoration quality in both line direction and details.

0.84dB, and 0.83dB at magniﬁcation factors ×2, ×3, and ×4, respectively, while using 46%, 52%, and 50% fewer parameters on the Manga109 dataset. Furthermore, com- pared to recent channel attention methods such as CARN, the proposed CRAFT achieves improvements of 1.03dB, 0.79dB, and 0.71dB at magniﬁcation factors ×2, ×3, and ×4, respectively, with a 54%, 53%, and 52% reduction in the number of parameters on the Manga109 dataset. Re- garding transformer-based methods [25, 21, 44], the pro- posed CRAFT gains performance improvements of 0.34dB, 0.31dB, and 0.29dB, respectively, with a comparable num- ber of parameters under the magniﬁcation factor of ×3 on the Manga109 dataset.

**Qualitative Results.** We present a visual comparison ( ×4) in Figure 3 and analyze the results. Our proposed CRAFT model integrates the strengths of both CNN and transformer structures, leading to accurate line direction re- covery while preserving image details. To further investi- gate the performance, we compare the local attribution map (LAM) [11] between CRAFT and SwinIR, as shown in Fig- ure 4. LAM indicates the correlation between the signiﬁ- cance of each pixel in LR and the SR of the patch that is outlined with the red box. By leveraging a broader range of information, our model achieves improved results. Fur- thermore, we examine the diffusion index (DI), which sig- niﬁes the range of pixels involved. A larger DI indicates a

wider scope of attention. Compared to SwinIR, our model exhibits a higher DI, implying that it can capture more con- textual information. These results demonstrate the effec- tiveness of the proposed CRAFT method.

**5.4. Ablation study**

**5.4.1 Effectiveness of HFERB and SRWAB**

We conduct several experiments to show the effectiveness of HFERB and SRWAB in Table 2. Speciﬁcally, we re- moved SRWAB and HFERB separately to assess their con- tributions. We observed that using local or global infor- mation alone, as in CRAFTconv and CRAFTtransformer , respectively, is insufﬁcient to learn a better representation (lower performance). Furthermore, we found that SR- WAB provides the most signiﬁcant performance improve- ment, demonstrating the beneﬁts of the long-range depen- dencies learned by the transformer. In addition, high- frequency priors from CNN are also helpful in restoring details, cross-reﬁning learned features and further improv- ing performance. Meanwhile, we also analyzed the prop- erties of HFERB and SRWAB from a frequency perspec- tive. We visualized the features extracted from two blocks in different RCRFGs and plotted the Fourier spectrum to observe what each block learns. The results, shown in Fig- ure 5, indicate that HFERB focuses more on high-frequency information, while SRWAB extracts more global informa-

|  |
| --- |
|  |

HR

|  |
| --- |
|  |

LR

|  |
| --- |
|  |

HR

|  |
| --- |
|  |

LR

SwinIR

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

DI: 10.78

SwinIR

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

DI: 7.36

CRAFT

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

DI: 30.34

CRAFT

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

DI: 18.89

|  |
| --- |
|  |

HR

|  |
| --- |
|  |

LR

|  |
| --- |
|  |

HR

|  |
| --- |
|  |

LR

SwinIR

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

DI: 9.37

SwinIR

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

DI: 9.30

CRAFT

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

DI: 30.94

CRAFT

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

DI: 17.98

|  |
| --- |
|  |

HR

|  |
| --- |
|  |

LR

|  |
| --- |
|  |

HR

|  |
| --- |
|  |

LR

SwinIR

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

DI: 13.03

SwinIR

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

DI: 12.88

CRAFT

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

DI: 30.87

CRAFT

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

DI: 32.17

Figure 4. Comparison of the LAM results of SwinIR [21] and CRAFT. LAM indicates the correlation between the signiﬁcance of each pixel in LR and the SR of the patch that is outlined with the red box. CRAFT utilizes a broader range of information to obtain better performance. DI quantiﬁes the LAM results, CRAFT has a higher DI score, indicating its ability to capture more contextual information.

SR Contribution LAM SR Contribution LAM

Results Area Results Results Area Results

SR Contribution LAM SR Contribution LAM

Results Area Results Results Area Results

LAM

Results

Contribution

Area

SR

Results

LAM

Results

Contribution

Area

SR

Results

Table 2. Study of HFERB, SRWAB, and HFB on SISR. The results

(×4) are obtained from the Manga109 dataset.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | HFERB | SRWAB | HFB | Concat | PSNR |
| CRAFT conv  CRAFT tranformer  CRAFT concat  CRAFT | √  √  √ | √  √  √ | √  √  √ | √ | 30.79  31.12  30.92  31.18 |

tion. Speciﬁcally, the top row of each image indicates the Fourier spectrum of each block, and the bottom row indi- cates the feature maps of each block. The ﬁgure shows that SRWAB has a weaker response and focuses more on the low-frequency parts, which correspond to ﬂat regions, while HFERB shows a stronger response and focuses more on in- tricate parts of features, such as edges and corners. The feature maps on the bottom row also support this conclu- sion. HFERB captures more details such as window edges and cornices, while SRWAB pays more attention to ﬂat ar- eas such as windows and walls.

**5.4.2 Effectiveness of HFB**

To evaluate the effectiveness of HFB, we conducted an ex- periment where we modiﬁed the fusion method to a con- catenation formulation. This involved concatenating the

Table 3. Effectiveness of high-frequency prior. The results (×4) are obtained from the Manga109 dataset.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Regular | Swap | Cascade | PSNR | SSIM |
| CRAFTswap  CRAFTcascade  CRAFT | √ | √ | √ | 30.67  30.88  31.18 | 0.9113  0.9141  0.9168 |

Table 4. Complexity analysis compared to SwinIR.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | #Params. | #FLOPs | #GPU Mem. | Ave. Time |
| (K) | (G) | (M) | (ms) |
| SwinIR | 897 | 32.2 | 141.2 | 72.0 |
| CRAFT | 753 | 26.1 | 79.5 | 42.8 |

HFERB and SRWAB output and replacing the HFB with a 3 × 3 convolutional layer to obtain the ﬁnal output. The re- sults are presented in Table 2, where CRAFTconcat denotes the modiﬁed version. The result shows that our proposed method outperforms the concatenation structure by 0.26dB, demonstrating the effectiveness of our HFB. The observed result can be attributed to SRWAB and HFERB focusing on disparate frequency information. Stacking features directly impedes the ability of the network to learn the relation- ship between high-frequency and low-frequency compo- nents. Conversely, the inter-attention mechanism presents a viable solution for integrating features with different dis- tributions.



LR Image



LR Image



SRWin-Attention

HFE

LFE

(a) Visualization of the spectrum of the 2nd RCRFG



SRWin-Attention

HFE

LFE

(b) Visualization of the feature of the 2nd RCRFG



SRWin-Attention

HFE

LFE

(e) Visualization of the spectrum of the 2nd RCRFG



SRWin-Attention

HFE

LFE

(f) Visualization of the feature of the 2nd RCRFG



SRWin-Attention

HFE

LFE

(c) Visualization of the spectrum of the 4th RCRFG



SRWin-Attention

HFE

LFE

(d) Visualization of the feature of the 4th RCRFG



SRWin-Attention

HFE

LFE

(g) Visualization of the spectrum of the 4th RCRFG



SRWin-Attention

HFE

LFE

(h) Visualization of the feature of the 4th RCRFG

Figure 5. Visualization of HFERB and SRWAB. The LFE indicates the local feature extraction branch in HFERB, the HFE means the

high-frequency enhancement branch in HFERB, and the SRWin-Attention represents the self-attention part in SRWAB.

**5.4.3 Effectiveness of High-Frequency Prior**

We conducted several experiments to investigate the effec- tiveness of high-frequency prior. Firstly, we swapped the input of **Q** and **K**, **V** in HFB and treated the output of SRWAB as **Q** and the output of HFERB as **K**, **V** to ver- ify whether global features are dominant in restoration and high-frequency features only serve as a prior for reﬁning the global representation. As shown in Table 3, compared to the original design, swapping the input leads to a signiﬁ- cant drop in performance, with a 0.51dB decrease in PSNR. Furthermore, we also performed an experiment to formulate the model as a cascade structure to verify the effectiveness of the design introducing high-frequency priors. As shown in Table 3, the CRAFTcascade structure resulted in a per- formance drop, with a 0.3dB decrease in PSNR compared to CRAFT. These results demonstrate the effectiveness of high-frequency priors in the CRAFT model.

**5.4.4 Complexity analysis**

We compared CRAFT with SwinIR in terms of complex- ity using an input size of 128 × 128, as shown in Table 4. The analysis considered parameters, FLOPs, GPU memory consumption, and average inference time. GPU memory was measured using the ofﬁcial PyTorch function, and time cost was calculated based on 100 inference runs. Com- pared to SwinIR, CRAFT has fewer parameters and FLOPs, and requires less memory consumption and inference time. Additionally, we analyzed the complexity of our CRAFT framework and summarized the ﬁndings in Table 5. We ob- served that SRWAB contributes approximately 46% of the total complexity, while HFERB involves fewer convolution operations, resulting in reduced FLOPs. Furthermore, the HFB module’s channel-wise attention effectively reduces

Table 5. Complexity analysis of each block.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | CRAFT  w/o HFERB | CRAFT  w/o SRWAB | CRAFT  w/o HFB | CRAFT |
| #Params. (K) | 688 | 441 | 503 | 753 |
| #FLOPs (G) | 23.8 | 14.2 | 20.0 | 26.1 |

the computational burden.

**6. Conclusion**

This paper investigates the impact of frequency on the performance of CNN and transformer structures in SISR and ﬁnds that transformer structures are more adept at cap- turing low-frequency information, but have limited capabil- ity to reconstruct high-frequency representations compared to CNN. To address this issue, we design a feature modu- lation transformer, named cross-reﬁnement adaptive feature modulation transformer (CRAFT), which comprises three key components: the high-frequency enhancement residual block (HFERB), the shift rectangle window attention block (SRWAB), and the hybrid fusion block (HFB). The HFERB is designed to extract high-frequency features, while the SRWAB captures global representations. In the HFB, we treat the output of HFERB as a high-frequency prior and the output of SRWAB as key and value, and use inter- attention to reﬁne the global representation. Experimental results demonstrate that CRAFT outperforms state-of-the- art methods by up to 0.29dB with relatively fewer parame- ters.

**Acknowledgements**

This work is supported by the National Natural Science Foundation of China (NSFC) under Grant U19A2052 and Grant 62020106011, as well as by the National Key Lab of Autonomous Intelligent Unmanned Systems.

参考文献

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Pro- ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, pages 126– 135, 2017. 6

[2] Namhyuk Ahn, Byungkon Kang, and Kyung Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *Proceedings of the European Confer- ence on Computer Vision (ECCV)*, pages 256–272, 2018. 2, 6, 7

[3] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 1–10, 2012. 6

[4] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yip- ing Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-Trained Image Processing Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vi- sion and Pattern Recognition (CVPR)*, pages 12299–12310, 2021. 1, 2

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

[6] Zheng Chen, Yulun Zhang, Jinjin Gu, Yongbing Zhang, and Linghe Kong. Cross aggregation transformer for image restoration. In *Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS)*, 2022. 1, 2, 5

[7] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional net- works. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2016. 1, 2

[8] Faming Fang, Juncheng Li, and Tieyong Zeng. Soft-edge assisted network for single image super-resolution. *IEEE Transactions on Image Processing*, 29:4656–4668, 2020. 1

[9] Guangwei Gao, Zhengxue Wang, Juncheng Li, Wenjie Li, Yi Yu, and Tieyong Zeng. Lightweight bimodal network for single-image super-resolution via symmetric cnn and recur- sive transformer. *Proceedings of the International Joint Con- ference on Artiﬁcial Intelligence (IJCAI)*, 2022. 6, 7

[10] Hayit Greenspan. Super-resolution in medical imaging. *The Computer Journal*, 52(1):43–63, 2009. 1

[11] 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 (CVPR)*, pages 9199–9208, 2021. 7

[12] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Sin- gle image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5197–5206, 2015. 6

[13] Zheng Hui, Yunchu Yang, Xinbo Gao, and Xiumei Wang.

Lightweight image super-resolution with information multi-

distillation network. In *Proceedings of the ACM Interna- tional Conference on Multimedia (ACMMM)*, pages 2024– 2032, 2019. 2, 6

[14] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal co- variate shift. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 448–456, 2015. 2

[15] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accu- rate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1646–1654, 2016. 2

[16] Christian Ledig, Lucas Theis, Ferenc Huszr, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo- realistic single image super-resolution using a generative ad- versarial network. In *Proceedings of the IEEE/CVF Confer- ence on Computer Vision and Pattern Recognition (CVPR)*, pages 4681–4690, 2017. 1

[17] Christian Ledig, Lucas Theis, Ferenc Huszr, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wen- zhe Shi. Photo-realistic single image super-resolution us- ing a generative adversarial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114, 2017. 1

[18] Wenbo Li, Xin Lu, Jiangbo Lu, Xiangyu Zhang, and Jiaya Jia. On efﬁcient transformer and image pre-training for low- level vision. *arXiv preprint arXiv:2112.10175*, 2021. 1, 2

[19] Wenbo Li, Kun Zhou, Lu Qi, Nianjuan Jiang, Jiangbo Lu, and Jiaya Jia. Lapar: Linearly-assembled pixel-adaptive re- gression network for single image super-resolution and be- yond. In *Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS)*, pages 1– 13, 2020. 2, 6

[20] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwang- gil Jeon, and Wei Wu. Feedback network for image super- resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3867–3876, 2019. 1

[21] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision Workshop (ICCVW)*, pages 1833–1844, 2021. 1, 2, 3, 6, 7, 8

[22] Jingyun Liang and Luc Van Gool. Flow-based Kernel Prior with Application to Blind Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pat- tern Recognition (CVPR)*, pages 10601–10610, 2021. 1

[23] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for sin- gle image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, pages 1132–1140, 2017. 2, 6, 7

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

*Proceedings of the IEEE/CVF Conference on Computer Vi- sion (ICCV)*, pages 10012–10022, 2021. 2

[25] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Lin- lin Zhang, and Tieyong Zeng. Transformer for single image super-resolution. In *Proceedings of the IEEE/CVF Confer- ence on Computer Vision and Pattern Recognition Workshop (CVPRW)*, pages 457–466, 2022. 1, 2, 6, 7

[26] Xiaotong Luo, Yuan Xie, Yulun Zhang, Yanyun Qu, Cui- hua Li, and Yun Fu. Latticenet: Towards lightweight image super-resolution with lattice block. In *Proceedings of the Eu- ropean Conference on Computer Vision (ECCV)*, pages 272– 289, 2020. 6

[27] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the IEEE/CVF Conference on Computer Vision (ICCV)*, pages 416–423, 2001. 6

[28] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Mul- timedia Tools and Applications*, 76(20):21811–21838, 2017. 6

[29] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image Super- Resolution with Non-Local Sparse Attention. In *Proceed- ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3517–3526, 2021. 1

[30] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S. Huang, and Humphrey Shi. Image super- resolution with cross-scale non-local attention and exhaus- tive self-exemplars mining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[31] Sivaram Prasad Mudunuri and Soma Biswas. Low resolu- tion face recognition across variations in pose and illumina- tion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(5):1034–1040, 2015. 1

[32] Kamyar Nazeri, Harrish Thasarathan, and Mehran Ebrahimi.

Edge-informed single image super-resolution. In *Proceed- ings of the IEEE/CVF Conference on Computer Vision Work- shop (ICCVW)*, pages 3275–3284, 2019. 1

[33] Wenzhe Shi, Jose Caballero, Ferenc Huszr, Johannes Totz, Andrew Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efﬁcient sub-pixel convolutional neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883, 2016. 4

[34] Bin Sun, Yulun Zhang, Songyao Jiang, and Yun Fu. Hy- brid pixel-unshufﬂed network for lightweight image super- resolution. In *Proceedings of the Association for the Ad- vancement of Artiﬁcial Intelligence (AAAI)*, 2023. 2, 6

[35] Radu Timofte, Vincent De Smet, and Luc Van Gool.

Anchored neighborhood regression for fast example-based super-resolution. In *Proceedings of the IEEE/CVF Confer- ence on Computer Vision (ICCV)*, pages 1920–1927, 2013. 1

[36] Radu Timofte, Vincent De Smet, and Luc Van Gool. A+: Adjusted anchored neighborhood regression for fast super- resolution. In *Proceedings of the Asian Conference on Com- puter Vision (ACCV)*, pages 111–126, 2014. 1

[37] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image Super-Resolution Using Dense Skip Connections. In *Pro- ceedings of the IEEE/CVF Conference on Computer Vision (ICCV)*, pages 4809–4817, 2017. 1

[38] Wenxiao Wang, Lu Yao, Long Chen, Binbin Lin, Deng Cai, Xiaofei He, and Wei Liu. Crossformer: A versatile vision transformer hinging on cross-scale attention. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. 5

[39] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Si- moncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Process- ing*, 13(4):600–612, 2004. 6

[40] Bin Xia, Yucheng Hang, Yapeng Tian, Wenming Yang, Qingmin Liao, and Jie Zhou. Efﬁcient non-local contrastive attention for image super-resolution. In *Proceedings of the Association for the Advancement of Artiﬁcial Intelligence (AAAI)*, pages 2759–2767, 2022. 1

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

[42] Syed Waqas Zamir, Aditya Arora, Salman Khan, Mu- nawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efﬁcient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5718–5729, 2022. 5, 6

[43] Roman Zeyde, Michael Elad, and Matan Protter. On sin- gle image scale-up using sparse-representations. In *Inter- national Conference on Curves and Surfaces (ICCS)*, pages 711–730, 2010. 6

[44] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang.

Efﬁcient long-range attention network for image super- resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2, 6, 7

[45] Xindong Zhang, Hui Zeng, and Lei Zhang. Edge-oriented convolution block for real-time super resolution on mobile devices. In *Proceedings of the ACM International Confer- ence on Multimedia (ACMMM)*, pages 4034–4043, 2021. 1

[46] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 294–310, 2018. 1

[47] Zhengdong Zhang and Vivienne Sze. Fast: A framework to accelerate super-resolution processing on compressed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1015–1024, 2017. 1