CoDe: An Explicit Content Decoupling Framework for Image Restoration

Enxuan Gu[1], Hongwei Ge[1]*, Yong Guo[2]
[1]Dalian University of Technology, [2]South China University of Technology

Abstract

*The performance of image restoration (IR) is highly dependent on the reconstruction quality of diverse contents with varying complexity. However, most IR approaches model the mapping between various complexity contents of inputs and outputs through the repeated feature calculation propagation mechanism in a unified pipeline, which leads to unsatisfactory results. To address this issue, we propose an explicit **Co**ntent **De**coupling framework for IR, dubbed **CoDe**, to end-to-end model the restoration process by utilizing decoupled content components in a divide-and-conquer-like architecture. Specifically, a Content Decoupling Module is first designed to decouple content components of inputs and outputs according to the frequency spectra adaptively generated from the transform domain. In addition, in order to harness the divide-and-conquer strategy for reconstructing decoupled content components, we propose an IR Network Container. It contains an optimized version, which is a streamlining of an arbitrary IR network, comprising the cascaded modulated subnets and a Reconstruction Layers Pool. Finally, a Content Consistency Loss is designed from the transform domain perspective to supervise the restoration process of each content component and further guide the feature fusion process. Extensive experiments on several IR tasks, such as image super-resolution, image denoising, and image blurring, covering both real and synthetic settings, demonstrate that the proposed paradigm can effectively take the performance of the original network to a new state-of-the-art level in multiple benchmark datasets (*e.g.*, **0.34**dB@Set5 $\times 4$ over DAT).*

1. Introduction

Image restoration (IR) has been a long-standing problem for its highly practical value in various low-level vision applications, such as image super-resolution (SR), denoising, and deblurring. It is a typical ill-posed problem due to the irreversible nature of the image degradation process. Recent advances in IR have been led by deep learning-based methods, as they can learn strong and generalizable priors from large-scale datasets. Existing deep learning-based IR approaches typically follow the two main architectures: convolutional neural networks (CNN) architecture and Transformer [38] architecture. For example, stacked denoising auto-encoder [39] is one of the best well-known CNN-based models for image denoising, Dong *et al*. proposed SRCNN [13] for image SR and multi-scale CNN (MSCNN) was proposed for image deblurring by Nah *et al*. [31]. The deep CNN models with the convolution operation have been proven to be effective for IR tasks [22, 28, 42, 52, 53]. As an alternative to CNN, Transformer-based [38] methods were designed via utilizing the self-attention mechanism to capture global interactions between contexts and has shown promising performance in IR problems [4, 6, 27]. Due to the powerful fitting ability of deep learning-based methods with the image-to-image regression strategy, the performance and reconstruction quality of IR has been improved significantly.

*Corresponding author. Email: hwge@dlut.edu.cn

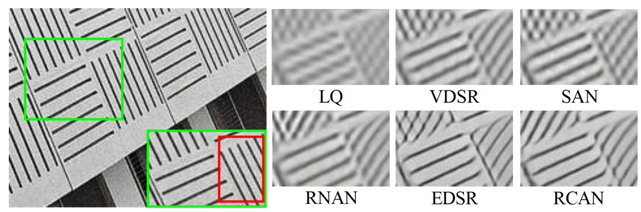

Figure 1. The most representative sample to illustrate our motivation: We observed that the contents in different regions of an image may follow different patterns/distributions, while many IR methods employ computationally complex networks consisting of repeated computational modules in a unified pipeline, to recover such regions resulting in wrongly restored results. This patch is img092 from Urban100 [18] dataset.

However, we have identified several issues with current CNN-based and Transformer-based IR methods. These methods have demonstrated excellent average restoration performance across various benchmarks, while they often occasionally produce poor/wrong results in some regions and obtain unsatisfactory results. As shown in Fig. 1, all the considered methods produce some anti-diagonal lines (from bottom left to top right) but the lines in the ground-truth image should be in the opposite direction, i.e., the diagonal (from top left to the bottom right). Overall, the quality of different regions in the restored/predicted image

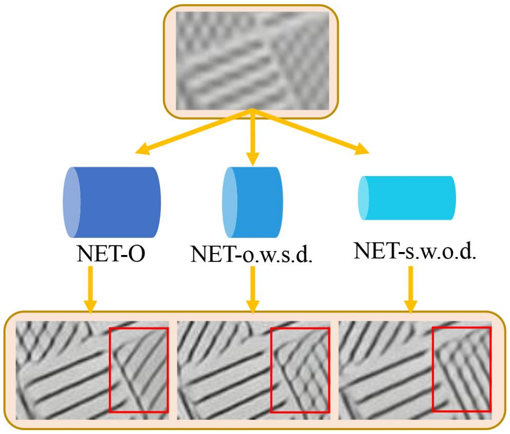

Figure 2. The alleviation of this overfitting phenomenon. This patch is fed to the original network (-O) and its simplified version with reduced computational complexity, respectively. The visually unsatisfactory restoration results are significantly mitigated, especially in the red box. NET stands for EDSR [28]. We modulate the computational cost by simply reducing the width and depth of the feature layers in the original network. 'Original Width, Smaller Depth' and 'Smaller Width, Original Depth' are abbreviated as -o.w.s.d and -s.w.o.d.

may vary a lot. This shortcoming will result in a significant decline in the performance when restoring such particular patches, which hinders the overall reconstruction ability. We hypothesize that different regions of an image should be processed in different ways instead of being treated equally. Here, we conduct a simple experiment to verify this. As shown in Fig. 2, we attribute the essence of recovering these visually unappealing observations to a kind of overfitting phenomenon, i.e., enforcing the model to process different contents/regions in a similar way, yielding a lack of diversity w.r.t. processing different contents. In other words, they treat the whole low-quality (LQ) image as the input, neglecting the characteristic that different contents/regions have varying patterns in the restoration process so that all the contents can only be treated equally in a unified pipeline through the repeated feature extraction layers. This behavior often results in some wrongly restored content that should be processed in a different way from the other regions. Surprisingly, we find that we can effectively correct these errors if we build a smaller network to reduce the overfitting risk. As shown in Fig. 2, either reducing the width or depth of the network successfully rectifies the errors. This result further reveals the necessity of processing different content components by modulating the original network to appropriate sub-nets with corresponding computational costs for IR tasks, and the content decoupling operation should be taken into consideration.

The aforementioned discussion leads to a series of research questions about the IR tasks naturally: 1) how to decouple the contents more rationally; 2) how to process decoupled content components in a more reasonable way for different IR tasks; 3) and how to supervise the whole restoration process under the new paradigm. We will try to answer these questions in Sec. 3.1, Sec. 3.2, and Sec. 3.3 respectively. To address the questions mentioned above, we propose an explicit content decoupling framework for image restoration. Specifically, we designed a Content Decoupling Module to disentangle the content components of LQ and HQ. This is achieved by adaptively generating frequency masks from the Discrete Cosine Transform (DCT) domain, and then using them to guide the content decoupling process. Besides, an IR Network Container is proposed to handle different content components in a divide-and-conquer manner, with each sub-net dedicated to handling a specific content component. To constrain the whole restoration process, we further design a Content Consistency Loss from the transform domain perspective, which forces the paradigm to pay more attention to the image contents with more complex patterns. We apply the proposed framework to the most current state-of-the-art IR methods. Experimental results demonstrate that our framework can effectively take the performance of the original network to a new state-of-the-art level.

The contributions of this work are as follows:

- We systematically and comprehensively answer the general questions about the image restoration tasks and propose a solution. The exploration of solving these issues can guide us to re-examine the challenges faced by image restoration from a different perspective.
- We propose an explicit **Co**ntent **De**coupling framework for image restoration, dubbed **CoDe**, that is needed to specify neither network architecture nor tasks, which can be applied to any existing image restoration network.
- Extensive experiments have been conducted on several real and synthetic IR tasks, demonstrating that our proposed framework can take the original network to a new state-of-the-art level in both performance and visual quality while maintaining attractive computational costs.

2. Related Work

Image Restoration. Existing deep learning-based IR methods can be divided into two main categories: CNN-based and Transformer-based methods. For the former, Dong *et al*. proposed SRCNN [13] for image SR, Zhang *et al*. proposed DnCNN [51] for image denoising, and Xu *et al*. proposed DCNN [44] for image deblurring. With the emergence of these pioneering work, a flurry of CNN-based methods, such as [10, 37, 41, 49, 55, 57], have been proposed to improve the representation ability for IR tasks. For the latter, after Chen *et al*. proposed a backbone model IPT [6] for various restoration problems based on the standard Transformer, there emerge a number of Transformer-based methods [7, 23, 25, 27, 48]. However, most of these

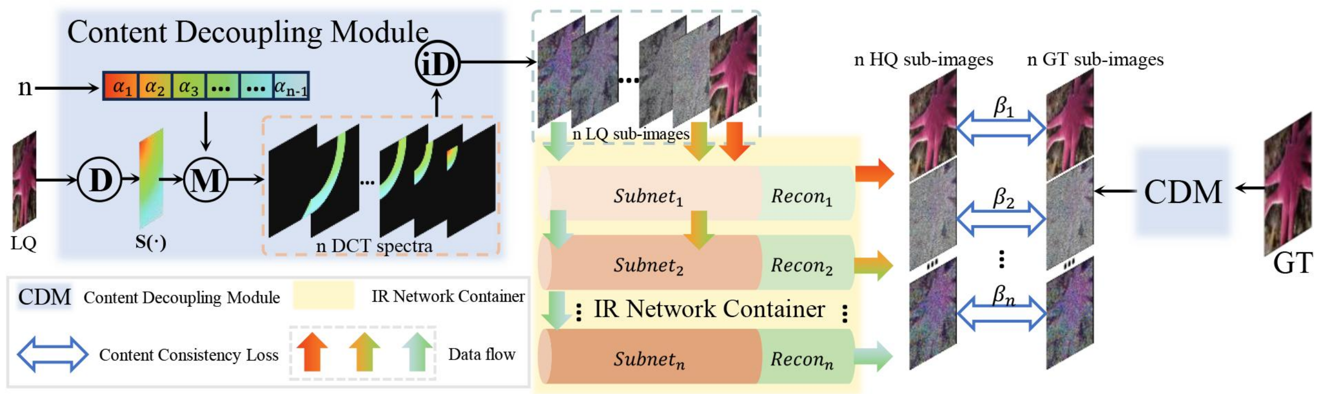

Figure 3. Overview of the proposed framework. Please refer to the color version for better view.

methods result in over-smoothed outputs and some textural details also fail to be recovered correctly. This is mainly because they neglect the difficulty of restoring diverse image contents can vary, and treat them indiscriminately.

Frequency Decoupling. Recently, researchers have proposed some deep learning-based IR methods, taking the characteristics of different frequency signals into account for content decoupling purposes. Li *et al*. [24] proposed a frequency decomposition operation for image SR to compensate for the lost information in the LQ image by large stride convolution. Magid *et al*. utilized a dynamic high-pass filter [29] that locally applies adaptive filter weights for each spatial location and channel group to preserve high-frequency signals. Fritsche *et al*. proposed the DSGAN [16] to generate LR-HR pairs via the low- and high-pass filters. Jiang *et al*. [20] and Zhang *et al*. [58] introduced the Octave Convolution (OctConv) [8] to obtain the low- and high-frequency information. There are also methods that use off-the-shelf operators for frequency decoupling. In [61], Zuo *et al*. introduced vertical and horizontal Sobel [14] operators to generate the gradient map for feature enhancement. In [26, 46, 60], Haar wavelet [12] was combined in the network to decompose images into 4 diverse content components. Shao *et al*. [34] and Chakrabarti *et al*. [5] adopted the Fast Fourier Transform (FFT) [11] to learn the global low- and high-frequency content components by converting the features to the frequency domain. However, using these frequency decoupling methods to achieve content decoupling has obvious drawbacks. Whether using the traditional convolution mechanism or applying off-the-shelf operators for content decoupling, they are essentially a weighted average operation, considering the weighted sum of pixel values within a fixed-size sliding window. Using such filters for all image patches to decouple content components may not be the best choice. In addition, the decomposition patterns are inflexible because one operation can only generate one subband and they can eventually decouple the content into only 2 or 4 subbands.

DCT-based IR Methods. DCT is widely used in traditional digital image processing, especially in JPEG image compression [40]. Unlike the aforementioned frequency decoupling methods, the DCT-based IR methods utilize all DCT coefficients, which contain information about all content components of the whole input. Refs [17, 45] use 8×8 block-wise DCT for image restoration, similar to JPEG image compression. However, this transformation will lead to severe visual discontinuity, resulting in poor visual quality. Although this problem can be alleviated by patch overlapping, it would introduce extra computational burdens.

3. Methodology

Overview. The contents in different regions of an image follow different patterns/distributions. Some regions are very smooth while others may contain a lot of high-frequency textures. In this case, treating them equally may compromise the performance in both regions since it is non-trivial to find a good trade-off between low-/high-frequency reconstruction. This directly inspires us to conduct content decoupling. To this end, we dive into the frequency domain for help and propose our explicit **Co**ntent **De**coupling framework, dubbed **CoDe**, for image restoration. It can decouple image contents from all the frequency signals and restore the decoupled content components respectively. Fig. 3 shows the overall architecture of the proposed framework. Our approach consists of three main components, Content Decoupling Module (CDM), IR Network Container (IRNC), and Content Consistency Loss (CCLoss), which are targeted to solve the three problems presented in Section 1. We will describe the technical details of these three components in the following subsections.

3.1. Content decoupling module

As the 8×8 block-wise DCT transformation would lead to visual discontinuity, we deprecate it and use the typical type-II DCT and type-II inverse DCT (iDCT) on the whole image for the content decoupling purpose. We denote these two operations by Ⓓ and ⓘⒹ in Fig. 3 respectively. The type-II DCT and iDCT formula can be expressed as the following Eq. 1 and Eq. 2:

$$S(u,v)=\frac{2}{\sqrt{W}\cdot\sqrt{H}}\times\sum_{x=0}^{W-1}\sum_{y=0}^{H-1}LQ(x,y)\cdot\cos\frac{(2x+1)u\pi}{2W}\cdot\cos\frac{(2y+1)v\pi}{2H}, \quad (1)$$

$$LQ_i(x,y)=\frac{2}{\sqrt{W}\cdot\sqrt{H}}\times\sum_{u=0}^{W-1}\sum_{v=0}^{H-1}S_i(u,v)\cos\frac{(2x+1)u\pi}{2W}\cos\frac{(2y+1)v\pi}{2H}, \quad (2)$$

where $LQ(\cdot)$ and $S(\cdot)$ represent the LQ in RGB color space and the whole DCT spectrum respectively; S_i and LQ_i, $i\in\{1,2,\ldots,n\}$ represent a collection of DCT spectrum and LQ sub-images with only one content component; x,y and u,v stand for the horizontal and vertical pixel index in RGB color space and in DCT spectrum; H, W denote the height and width of the image. Here, we have:

$$\begin{cases}LQ(\cdot)=LQ_1\oplus LQ_2\oplus\ldots\oplus LQ_n,\\ S(\cdot)=S_1\oplus S_2\oplus\ldots\oplus S_n,\end{cases} \quad (3)$$

where $\oplus$ denotes the pixel-wise summation. Note that we utilize the type-II DCT on the whole image can ensure all pixel information contributes to the content decoupling process. This operation is equivalent to increasing the receptive field of the model in a disguised form, which allows the model to perceive the content information contained in each pixel.

The specific mechanism of the CDM is as follows. Taking LQ as an example, we first set a hyperparameter n, representing the number of content components to be decoupled. Then n–1 numbers different from each other between (0,1) are randomly initialized and sorted in ascending order, $[\alpha_1,\alpha_2,...,\alpha_{n-1}]$, which are used to determine the radius of the frequency mask $[r_0,r_1,...,r_{n-1}]$, where

$$r_i=\begin{cases}\alpha_i*\sqrt{H^2+W^2} & ,i\in\{1,2,\ldots,n-1\},\\ 0 & ,i=0.\end{cases} \quad (4)$$

Regarding the pixel in the upper left corner of the DCT spectrum $S(\cdot)$ obtained by Eq. 1 as the center of the circle, r_{i-1} and r_i, $i\in\{1,2,\ldots,n-1\}$ as the inner and outer radii along the diagonal direction to separate $S(\cdot)$ into n subbands $[S_1,S_2,...,S_n]$. We call this operation Content Masking and denote it by Ⓜ. In this way, we get n DCT spectra, and for each S_i it only contains a specific range of frequency signal. Then performing the iDCT by Eq. 2 on $S_i, i\in\{1,2,\ldots,n\}$ to obtain a collection of LQ sub-images, $[LQ_1,LQ_2,...,LQ_n]$. Each LQ_i only contains one frequency component as well. Similarly, after applying the above operation to ground truth (GT), we obtain a collection of GT sub-images, $[GT_1,GT_2,...,GT_n]$, and we have:

$$GT(\cdot)=GT_1\oplus GT_2\oplus\ldots\oplus GT_n, \quad (5)$$

where $\oplus$ denotes the pixel-wise summation. Note that $[\alpha_1,\alpha_2,...,\alpha_{n-1}]$ are learnable and self-adaptive, and our method differs from all previous frequency decoupling ones because it can decouple the image content into arbitrary number of parts by changing the hyperparameter n.

3.2. IR network container

In this part, we elaborately design IRNC by employing the divide-and-conquer strategy. It mainly consists of a simplified version, *i.e*., a streamlining of an arbitrary IR network, which contains a collection of cascaded modulated sub-nets with different computation costs and corresponding reconstruction layers following after, which we call the Reconstruction Layer Pool (abbr. $Recon$). Specifically, when setting $n>1$, there will be n LQ sub-images, LQ_1, LQ_2, $\ldots$, LQ_n, generated from the CDM. Then, IRNC will transfer into an optimized version with n subnets, $Subnet_1$, $Subnet_2$, $\ldots$, $Subnet_n$, and cascading together. Subsequently, each LQ_i will be fed to corresponding $Subnet_i$ to obtain the reconstructed high-dimensional features. After that, $Recon$ receives these features and transforms them back to RGB space through the corresponding reconstruction layer $Recon_i(\cdot)$ to obtain n HQ sub-images $HQ_1, HQ_2, \ldots, HQ_n$. If setting n to 1, it denotes the content decoupling operation will be omitted and IRNC will be degraded to the original network. In other words, this setup is equivalent to restoring LQ by the original network, which we denote as $ORNet(\cdot)$. The whole process of IRNC can be described formally as follows:

$$\begin{cases}HQ=ORNet(LQ) & ,n=1,\\ HQ_i=Recon_i(Subnet_i(\cdots(Subnet_1(LQ_i)))) & ,i\in\{1,2,\ldots,n\}.\end{cases} \quad (6)$$

We build a set of subnets with different model sizes by adopting the parameter sharing technique [32] to process different decoupled contents. Without loss of generality, $Subnet_i$ with a smaller index has a lower computational cost and its model size gradually grows with the increase of index i. Note that the smaller the sub-net index number is, the fewer the number of feature layers and channels it contains. Such modulations will bring about a shift in computation cost, and a reduction in the number of parameters.

3.3. Content consistency loss

As our paradigm is multi-output, commonly used loss functions in image restoration are invalid, such as the L1 Loss and L2 Loss, and in order to maintain the consistency of diverse content components between HQ and GT, we design the CCLoss from the frequency domain perspective to replace the term corresponding to the pixel-wise reconstruction loss in the original method, which is represented formally as follows:

$$\mathcal{L}(\beta;\Theta)=\sum_{i=1}^{n}\beta_i\cdot\|GT_i-HQ_i\|_1, \quad (7)$$

Table 1. Quantitative comparison (PSNR/SSIM) for **classical image SR** with state-of-the-art methods on benchmark datasets.

Scale	Method	Years	TrainSet	Set5 [2]		Set14 [50]		BSD100 [30]		Urban100 [18]	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
×2	ELAN [55]	ECCV2022	DIV2K	38.36	0.9620	34.20	0.9228	32.45	0.9030	33.44	0.9391
	Ours			38.54	0.9642	34.43	0.9239	32.63	0.9039	33.50	0.9402
	Omni-SR [41]	CVPR2023	DIV2K	38.22	0.9613	33.98	0.9210	32.36	0.9020	33.05	0.9363
	Ours			38.34	0.9622	34.10	0.9219	32.44	0.9026	33.12	0.9369
	HAT [7]	CVPR2023	DF2K	38.58	0.9628	34.70	0.9261	32.59	0.9050	34.31	0.9459
	Ours			38.64	0.9632	34.78	0.9267	32.68	0.9058	34.42	0.9467
	SRFormer [59]	ICCV2023	DIV2K	38.23	0.9613	33.94	0.9209	32.36	0.9019	32.91	0.9353
	Ours			38.35	0.9622	34.08	0.9215	32.44	0.9022	33.02	0.9358
	DAT [9]	ICCV2023	DF2K	38.58	0.9629	34.81	0.9272	32.61	0.9051	34.37	0.9458
	Ours			38.72	0.9633	34.93	0.9279	32.71	0.9056	34.51	0.9472
	GRL [25]	CVPR2023	DIV2K	38.67	0.9647	35.08	0.9303	32.67	0.9087	35.06	0.9505
	Ours			**38.82**	**0.9656**	**35.17**	**0.9311**	**32.81**	**0.9092**	**35.12**	**0.9510**
×4	ELAN [55]	ECCV2022	DIV2K	32.75	0.9022	28.96	0.7914	27.83	0.7459	27.13	0.8167
	Ours			32.86	0.9031	29.08	0.7926	27.92	0.7467	27.34	0.8175
	Omni-SR [41]	CVPR2023	DIV2K	32.49	0.8988	28.78	0.7859	27.71	0.7415	26.64	0.8018
	Ours			32.57	0.8996	28.92	0.7868	27.79	0.7422	26.83	0.8023
	HAT [7]	CVPR2023	DF2K	32.92	0.9047	29.15	0.7958	27.97	0.7505	27.87	0.8346
	Ours			33.01	0.9054	29.22	0.7962	28.09	0.7509	27.93	0.8354
	SRFormer [59]	ICCV2023	DIV2K	32.51	0.8988	28.82	0.7872	27.73	0.7422	26.67	0.8032
	Ours			32.65	0.9001	28.98	0.7892	27.89	0.7469	26.81	0.8051
	DAT [9]	ICCV2023	DF2K	33.08	0.9055	29.23	0.7973	28.00	0.7515	27.87	0.8343
	Ours			**33.42**	0.9072	29.41	0.7982	28.13	0.7522	27.98	0.8349
	GRL [25]	CVPR2023	DIV2K	33.10	0.9094	29.37	0.8058	28.01	0.7611	28.53	0.8504
	Ours			33.21	**0.9101**	**29.46**	**0.8071**	**28.13**	**0.7619**	**28.61**	**0.8513**

where Θ denotes all the learnable parameters of our framework, and $\beta_i, i \in \{1, 2, \ldots, n\}$ denote the weighting coefficients of different content components. This loss function ensures our framework recovers high-quality images by minimizing the distance between the HQ sub-images and GT sub-images corresponding to different component components. Essentially, it minimizes the DCT coefficients between HQ sub-images and GT sub-images. Under the supervision of the CCLoss, the final restoration result can be formally expressed as:

$$HQ(\cdot) = HQ_1 \oplus HQ_2 \oplus \ldots \oplus HQ_n, \tag{8}$$

where $\oplus$ represents the pixel-wise summation.

4. Experimental Results and Analysis

The experimental results are shown in this section. We first report the quantitative and visual comparisons for image SR, image denoising, and image deblurring covering both real and synthetic scenes. The configurations of hyperparameters are as follows: for image SR, we set n to 3, and $\beta_1, \beta_2, \beta_3 = 0.3, 0.7, 1$; for image denoising, n is set to 2, and $\beta_1, \beta_2 = 0.3, 1$; for image deblurring, we set n to 3, and $\beta_1, \beta_2, \beta_3 = 0.3, 0.5, 1$ respectively. Note that, for fair comparisons, we retrain all the models on 4 Nvidia Tesla V100 GPUs using the code released by the original author, while the rest of the configurations remain consistent with the original work, such as the training datasets, input patch size, batch size, and the testing configurations. For each task, we apply commonly used datasets for testing and report PSNR (dB) and/or SSIM [43] to evaluate the performance of our method.

4.1. Results on image SR

Classical image SR. We first apply our framework to some most state-of-the-art SISR methods: DAT [9], SRFormer [59], ELAN [55], HAT [7], Omni-SR [41], and GRL [25]. Set5 [2], Set14 [50], BSD100 [30], and Urban100 [18] are used for evaluation. Both CNN-based models and Transformer-based SR models are summarized. Table 1 shows the quantitative comparisons. As observed, our method demonstrates performance improvements across all four benchmark datasets for each scale factor. The maximum increase in PSNR is **0.34** dB on Set5 [2] at ×2 scale for DAT [9]. It is important to note that our approach does not alter the network architecture of the original method but instead focuses on optimizing the computation flow by considering the content components of the input data. Visual comparisons are presented in the first 2 rows in Fig. 4, where we perform ×4 image SR using several most current

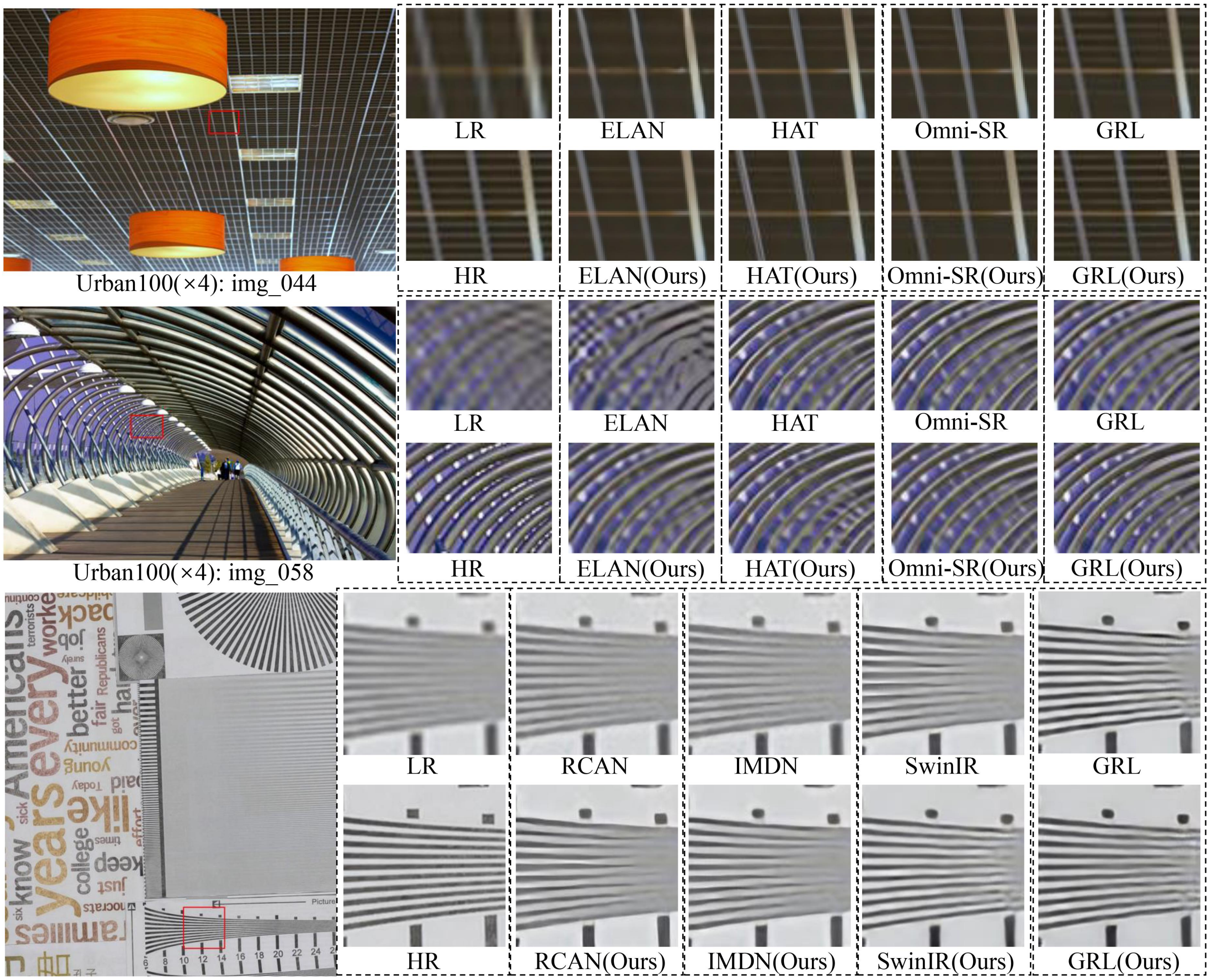

Figure 4. Visual comparisons of $\times 4$ **classical image SR** (first 2 rows) and **real-world image SR** (3rd row). These 3 patches are from Urban100 [18] and RealSR [3]. Best viewed by zooming.

Table 2. Quantitative comparison (average PSNR) with state-of-the-art methods for **color** image denoising results on benchmark datasets.

Method	CBSD68 [30]			Kodak24 [15]			McMaster [54]			Urban100 [18]		
	$\sigma=15$	$\sigma=25$	$\sigma=50$	$\sigma=15$	$\sigma=25$	$\sigma=50$	$\sigma=15$	$\sigma=25$	$\sigma=50$	$\sigma=15$	$\sigma=25$	$\sigma=50$
RNAN [57]	-	-	28.27	-	-	29.58	-	-	29.72	-	-	29.08
Ours	-	-	**28.40**	-	-	**29.63**	-	-	**29.89**	-	-	**29.52**
SwinIR [27]	34.42	31.78	28.56	35.34	32.89	29.79	35.61	33.20	30.22	35.13	32.90	29.82
Ours	**34.56**	**32.01**	**28.74**	**35.41**	**33.03**	**29.92**	**35.74**	**33.29**	**30.33**	**35.21**	**33.09**	**29.95**
DRUNet [52]	34.30	31.69	28.51	35.31	32.89	29.86	35.40	33.14	30.08	34.81	32.60	29.61
Ours	**34.41**	**31.76**	**28.57**	**35.43**	**32.97**	**29.92**	**35.51**	**33.20**	**30.13**	**34.90**	**32.67**	**29.65**
GRL [25]	34.45	31.82	28.62	35.43	33.02	29.93	35.73	33.46	30.36	35.54	33.35	30.46
Ours	**34.52**	**31.89**	**28.68**	**35.56**	**33.11**	**29.99**	**35.84**	**33.54**	**30.41**	**35.69**	**33.48**	**30.51**

state-of-the-art methods. As depicted, our framework effectively restores high-frequency details, reducing blurring artifacts and resulting in sharper and more natural edges.

Real-world image SR. The ultimate goal of image SR is for real-world applications. So, we also apply our framework to several state-of-the-art real-world image SR methods for comparison on the RealSR [3] dataset. The qualitative results are shown in the 3rd row in Fig. 4, and similar conclusions can be drawn from the visual comparisons. Our method produces visually pleasing images with clearer and sharper edges over the original methods both in CNN-based [19, 56] and Transformer-based [25, 27] real-world

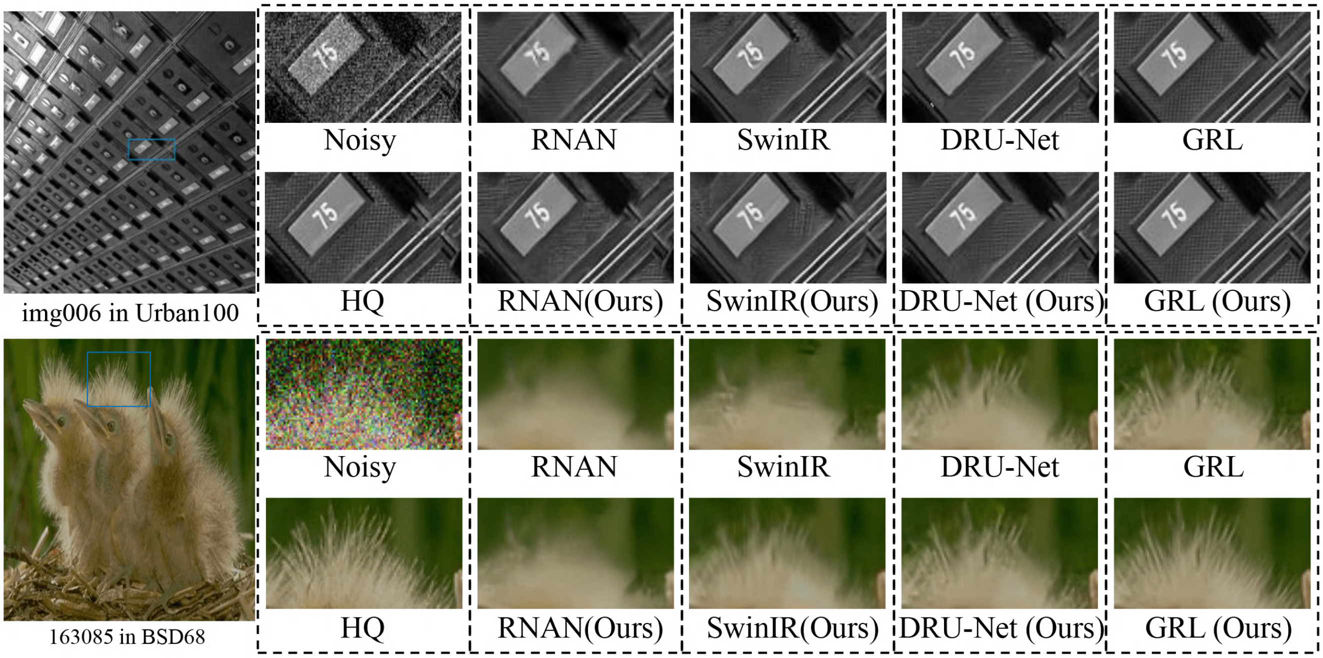

Figure 5. Visual comparison of **grayscale** and **color** Gaussian image denoising with noise level σ=50. Zoom in for more details.

SR models. This shows that our method can deal with more complex corruptions and achieves better performance in real-world scenes.

4.2. Results on image denoising

Besides the experiments on image SR, we also investigate the effectiveness of our framework on image denoising.

Table 3. Quantitative comparison (average PSNR) with state-of-the-art methods for **grayscale** image denoising results on benchmark datasets.

Method	Set12 [51]			BSD68 [30]			Urban100 [18]		
	σ=15	σ=25	σ=50	σ=15	σ=25	σ=50	σ=15	σ=25	σ=50
RNAN [57]	-	-	27.70	-	-	26.48	-	-	27.65
Ours	-	-	**27.83**	-	-	**26.55**	-	-	**27.73**
SwinIR [27]	33.36	31.01	27.91	31.97	29.50	26.58	33.70	31.30	27.98
Ours	**33.48**	**31.11**	**27.99**	**32.12**	**29.68**	**26.65**	**33.82**	**31.41**	**28.09**
DRUNet [52]	33.25	30.94	27.90	31.91	29.48	26.59	33.44	31.11	27.96
Ours	**33.34**	**31.02**	**27.96**	**32.06**	**29.60**	**26.65**	**33.57**	**31.23**	**28.02**
GRL [25]	33.47	31.12	28.03	32.00	29.54	26.60	34.09	31.80	28.59
Ours	**33.64**	**31.23**	**28.10**	**32.16**	**29.63**	**26.69**	**34.24**	**31.89**	**28.71**

Synthetic Image Denoising. First, the experimental results on color and grayscale Gaussian image denoising are shown in Table 2 and Table 3. We apply our framework to state-of-the-art denoising methods: RNAN [57], SwinIR [27], DRU-Net [52], and GRL [25]. Set12 [51], BSD68 [30], Urban100 [18], McMaster [54], and Kodak24 [15] are used for evaluation. Our framework can improve the performance of these methods across all datasets and noise levels. The performance gain can be attributed to the content decoupling strategy, which effectively captures content distribution throughout the degraded image, proving particularly effective for heavy noise levels. Notably, our method outperforms the state-of-the-art model GRL [25] by up to **0.15** dB on the large Urban100 [18] dataset at σ=15 noise level. We also show visual results for grayscale and color image denoising in Fig. 5. Under our framework, these methods can remove more noise and preserve more high-frequency image details, resulting in higher content fidelity.

Table 4. **Real-world denoising** comparisons on SIDD [1] and DND [33] datasets.

Method	Years	SIDD [1]		DND [33]	
		PSNR	SSIM	PSNR	SSIM
Cycle-ISP [47]	CVPR2020	39.52	0.957	39.56	0.956
Ours		**39.61**	**0.958**	**39.70**	**0.958**
AINDNet [21]	CVPR2020	38.95	0.952	39.37	0.951
Ours		**39.05**	**0.954**	**39.52**	**0.953**
NBNet [10]	CVPR2021	39.75	0.973	39.89	0.955
Ours		**39.84**	**0.974**	**39.98**	**0.957**
MIRNet-v2 [49]	T-PAMI2022	39.84	0.959	39.86	0.955
Ours		**39.95**	**0.962**	**39.99**	**0.957**

Real Image Denoising. We also demonstrate the effectiveness of the proposed framework for real image denoising, and apply it to state-of-the-art methods: CycleISP [47], AINDNet [21], NBNet [10], and MIRNet-v2 [49], and evaluate them on both SIDD [1] and DND [33] datasets. Quantitative comparisons in terms of PSNR and SSIM metrics are summarized in Table 4. When compared to the recent best methods, our algorithm makes a performance improvement of **0.11** dB over MIRNet-v2 [49] on SIDD and **0.15** dB over AINDNet [21] on DND, which indicates the generalization ability of our framework. Fig. 6 shows the visual compar-

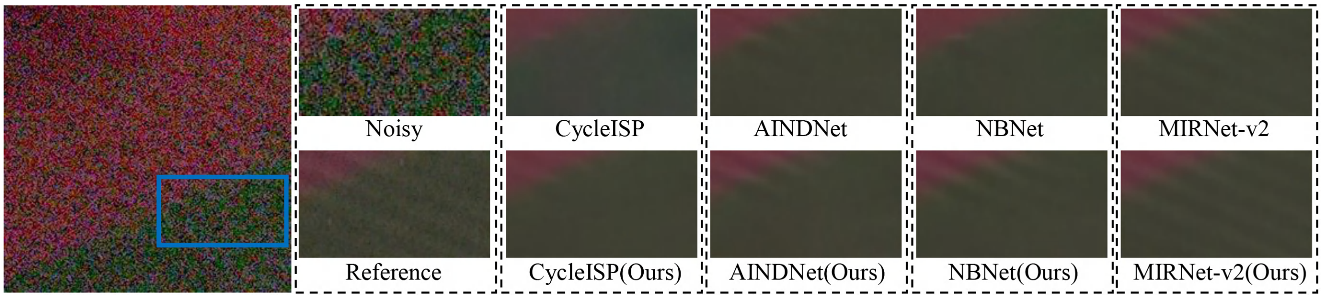

Figure 6. Visual comparisons of the denoising examples from SIDD [1] dataset with **real-world noises**. Zoom in for a better view.

Figure 7. Visual comparisons of the **single-image motion deblurring** on GoPro [31] (1st row) and HIDE [35] (2nd row). Zoom in for more details.

isons after our framework is applied to these state-of-the-art methods for real noise removal. The results show that with the proposed framework these methods are more effective in removing real noise and can produce more perceptually pleasing outputs.

Table 5. **Single-image motion deblurring** results on GoPro [31] and HIDE [35] datasets.

Method	Years	GoPro [31]		HIDE [35]	
		PSNR	SSIM	PSNR	SSIM
Stripformer [36]	ECCV2022	33.08	0.962	31.03	0.940
Ours		**33.19**	**0.963**	**31.15**	**0.942**
BANet [37]	TIP2022	32.54	0.957	30.16	0.930
Ours		**32.69**	**0.959**	**30.42**	**0.932**
GRL [25]	CVPR2023	33.93	0.968	31.65	0.947
Ours		**34.05**	**0.971**	**31.76**	**0.948**

4.3. Results on image deblurring

Single image motion deblurring. We further investigate the effectiveness of the proposed framework on image deblurring. We apply our approach to state-of-the-art methods: GRL [25], Stripformer [36], and BANet [37]. Table. 5 shows the experimental results on GoPro [31] and HIDE [35] datasets. As observed, when our framework is equipped with the most recent state-of-the-art methods GRL [25], the PSNR is significantly improved by **0.12** dB and **0.11** dB on GoPro [31] and HIDE [35] respectively. The visual comparisons are shown in Fig. 7. We can find that with our proposed framework, these methods can obtain more detailed structures and clearer contours than the original models, which also illustrates the feasibility and effectiveness of our framework.

5. Conclusion

In this paper, we propose a framework with content decoupling capacities for image restoration, dubbed CoDe, which can explicitly model the mapping between the diverse content patterns of the inputs and outputs through a divide-and-conquer-like architecture in an end-to-end manner. Comprehensive experiments across various image restoration tasks, including image super-resolution, image denoising, and image deblurring, conducted in both real-world and synthetic scenes, showcase the effectiveness of the proposed paradigm, and it can successfully elevate the performance of the original network to a new state-of-the-art level with a significant gain across multiple benchmark datasets.

Acknowledgements. This work is supported by the National Natural Science Foundation of China (61976034), the Dalian Science and Technology Innovation Fund (2022JJ12GX013), the Liaoning Natural Science Foundation (2022-YGJC20), and the Fundamental Research Funds for the Central Universities(DUT23YG103).

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1692–1700, 2018. 7, 8

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

[3] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3086–3095, 2019. 6

[4] Jiezhang Cao, Yawei Li, Kai Zhang, Jingyun Liang, and Luc Van Gool. Video super-resolution transformer. *arXiv preprint arXiv:2106.06847*, 2021. 1

[5] Ayan Chakrabarti. A neural approach to blind motion deblurring. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 221–235. Springer, 2016. 3

[6] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021. 1, 2

[7] Xiangyu Chen, Xintao Wang, Wenlong Zhang, Xiangtao Kong, Yu Qiao, Jiantao Zhou, and Chao Dong. Hat: Hybrid attention transformer for image restoration. *arXiv preprint arXiv:2309.05239*, 2023. 2, 5

[8] Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3435–3444, 2019. 3

[9] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xiaokang Yang, and Fisher Yu. Dual aggregation transformer for image super-resolution. In *ICCV*, 2023. 5

[10] Shen Cheng, Yuzhi Wang, Haibin Huang, Donghao Liu, Haoqiang Fan, and Shuaicheng Liu. Nbnet: Noise basis learning for image denoising with subspace projection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4896–4906, 2021. 2, 7

[11] James W Cooley and John W Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of computation*, 19(90):297–301, 1965. 3

[12] Ingrid Daubechies. The wavelet transform, time-frequency localization and signal analysis. *IEEE transactions on information theory*, 36(5):961–1005, 1990. 3

[13] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 1, 2

[14] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification and scene analysis*. Wiley New York, 1973. 3

[15] Rich Franzen. Kodak lossless true color image suite. *source: http://r0k. us/graphics/kodak*, 4(2), 1999. 6, 7

[16] Manuel Fritsche, Shuhang Gu, and Radu Timofte. Frequency separation for real-world super-resolution. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3599–3608. IEEE, 2019. 3

[17] Tiantong Guo, Hojjat Seyed Mousavi, and Vishal Monga. Adaptive transform domain image super-resolution via orthogonally regularized deep networks. *IEEE transactions on image processing*, 28(9):4685–4700, 2019. 3

[18] 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. 1, 5, 6, 7

[19] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *Proceedings of the 27th acm international conference on multimedia*, pages 2024–2032, 2019. 6

[20] Bo Jiang, Jinxing Li, Huafeng Li, Ruxian Li, David Zhang, and Guangming Lu. Enhanced frequency fusion network with dynamic hash attention for image denoising. *Information Fusion*, 92:420–434, 2023. 3

[21] Yoonsik Kim, Jae Woong Soh, Gu Yong Park, and Nam Ik Cho. Transfer learning from synthetic to real-noise denoising with adaptive instance normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3482–3492, 2020. 7

[22] Christian Ledig, Lucas Theis, Ferenc Huszár, 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 adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 1

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

[24] Xin Li, Xin Jin, Tao Yu, Simeng Sun, Yingxue Pang, Zhizheng Zhang, and Zhibo Chen. Learning omni-frequency region-adaptive representations for real image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1975–1983, 2021. 3

[25] Yawei Li, Yuchen Fan, Xiaoyu Xiang, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Efficient and explicit modelling of image hierarchies for image restoration. *arXiv preprint arXiv:2303.00748*, 2023. 2, 5, 6, 7, 8

[26] Zhen Li, Zeng-Sheng Kuang, Zuo-Liang Zhu, Hong-Peng Wang, and Xiu-Li Shao. Wavelet-based texture reformation network for image super-resolution. *IEEE Transactions on Image Processing*, 31:2647–2660, 2022. 3

[27] 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 international conference on computer vision*, pages 1833–1844, 2021. 1, 2, 6, 7

[28] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 1, 2

[29] Salma Abdel Magid, Yulun Zhang, Donglai Wei, Won-Dong Jang, Zudi Lin, Yun Fu, and Hanspeter Pfister. Dynamic high-pass filtering and multi-spectral attention for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4288–4297, 2021. 3

[30] 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 Eighth IEEE International Conference on Computer Vision. ICCV 2001*, pages 416–423. IEEE, 2001. 5, 6, 7

[31] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3883–3891, 2017. 1, 8

[32] Hieu Pham, Melody Guan, Barret Zoph, Quoc Le, and Jeff Dean. Efficient neural architecture search via parameters sharing. In *International conference on machine learning*, pages 4095–4104. PMLR, 2018. 4

[33] Tobias Plotz and Stefan Roth. Benchmarking denoising algorithms with real photographs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1586–1595, 2017. 7

[34] Mingwen Shao, Yuanjian Qiao, Deyu Meng, and Wangmeng Zuo. Uncertainty-guided hierarchical frequency domain transformer for image restoration. *Knowledge-Based Systems*, page 110306, 2023. 3

[35] Ziyi Shen, Wenguan Wang, Xiankai Lu, Jianbing Shen, Haibin Ling, Tingfa Xu, and Ling Shao. Human-aware motion deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5572–5581, 2019. 8

[36] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Stripformer: Strip transformer for fast image deblurring. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIX*, pages 146–162. Springer, 2022. 8

[37] Fu-Jen Tsai, Yan-Tsung Peng, Chung-Chi Tsai, Yen-Yu Lin, and Chia-Wen Lin. Banet: A blur-aware attention network for dynamic scene deblurring. *IEEE Transactions on Image Processing*, 31:6789–6799, 2022. 2, 8

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[39] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008. 1

[40] G.K. Wallace. The jpeg still picture compression standard. *IEEE Transactions on Consumer Electronics*, 38(1):xviii–xxxiv, 1992. 3

[41] Hang Wang, Xuanhong Chen, Bingbing Ni, Yutian Liu, and Jinfan Liu. Omni aggregation networks for lightweight image super-resolution. *arXiv preprint arXiv:2304.10244*, 2023. 2, 5

[42] 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 *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 1

[43] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[44] Li Xu, Jimmy S Ren, Ce Liu, and Jiaya Jia. Deep convolutional neural network for image deconvolution. *Advances in neural information processing systems*, 27, 2014. 2

[45] Ruyu Xu, Xuejing Kang, Chunxiao Li, Hong Chen, and Anlong Ming. Dct-fanet: Dct based frequency attention network for single image super-resolution. *Displays*, 74: 102220, 2022. 3

[46] Mengping Yang, Zhe Wang, Ziqiu Chi, and Wenyi Feng. Wavegan: Frequency-aware gan for high-fidelity few-shot image generation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 1–17. Springer, 2022. 3

[47] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Cycleisp: Real image restoration via improved data synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2696–2705, 2020. 7

[48] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022. 2

[49] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for fast image restoration and enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1934–1948, 2022. 2, 7

[50] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7*, pages 711–730. Springer, 2012. 5

[51] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017. 2, 7

[52] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pat-*

tern Analysis and Machine Intelligence, 44(10):6360–6376, 2021. 1, 6, 7

[53] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021. 1

[54] Lei Zhang, Xiaolin Wu, Antoni Buades, and Xin Li. Color demosaicking by local directional interpolation and nonlocal adaptive thresholding. *Journal of Electronic imaging*, 20(2): 023016–023016, 2011. 6, 7

[55] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 649–667. Springer, 2022. 2, 5

[56] 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 286–301, 2018. 6

[57] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. *arXiv preprint arXiv:1903.10082*, 2019. 2, 6, 7

[58] Yanni Zhang, Qiang Li, Miao Qi, Di Liu, Jun Kong, and Jianzhong Wang. Multi-scale frequency separation network for image deblurring. *arXiv preprint arXiv:2206.00798*, 2022. 3

[59] 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. 5

[60] Zhiyu Zhu, Junhui Hou, Jie Chen, Huanqiang Zeng, and Jiantao Zhou. Hyperspectral image super-resolution via deep progressive zero-centric residual learning. *IEEE Transactions on Image Processing*, 30:1423–1438, 2020. 3

[61] Yifan Zuo, Jiacheng Xie, Hao Wang, Yuming Fang, Deyang Liu, and Wenying Wen. Gradient-guided single image super-resolution based on joint trilateral feature filtering. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 3