Dual-domain Adaptation Networks for Realistic Image Super-resolution

Chaowei Fang, *Member, IEEE*, Bolin Fu, De Cheng, Lechao Cheng, Guanbin Li, *Member, IEEE*

***Abstract*—Realistic image super-resolution (SR) focuses on transforming real-world low-resolution (LR) images into high-resolution (HR) ones, handling more complex degradation patterns than synthetic SR tasks. This is critical for applications like surveillance, medical imaging, and consumer electronics. However, current methods struggle with limited real-world LR-HR data, impacting the learning of basic image features. Pre-trained SR models from large-scale synthetic datasets offer valuable prior knowledge, which can improve generalization, speed up training, and reduce the need for extensive real-world data in realistic SR tasks. In this paper, we introduce a novel approach, *Dual-domain Adaptation Networks*, which is able to efficiently adapt pre-trained image SR models from simulated to real-world datasets. To achieve this target, we first set up a spatial-domain adaptation strategy through selectively updating parameters of pre-trained models and employing the low-rank adaptation technique to adjust frozen parameters. Recognizing that image super-resolution involves recovering high-frequency components, we further integrate a frequency domain adaptation branch into the adapted model, which combines the spectral data of the input and the spatial-domain backbone's intermediate features to infer HR frequency maps, enhancing the SR result. Experimental evaluations on public realistic image SR benchmarks, including RealSR, D2CRealSR, and DRealSR, demonstrate the superiority of our proposed method over existing state-of-the-art models. Codes are available at: https://github.com/dummerchen/DAN.**

***Index Terms*—Image super-resolution, neural network adaptation, spectral data.**

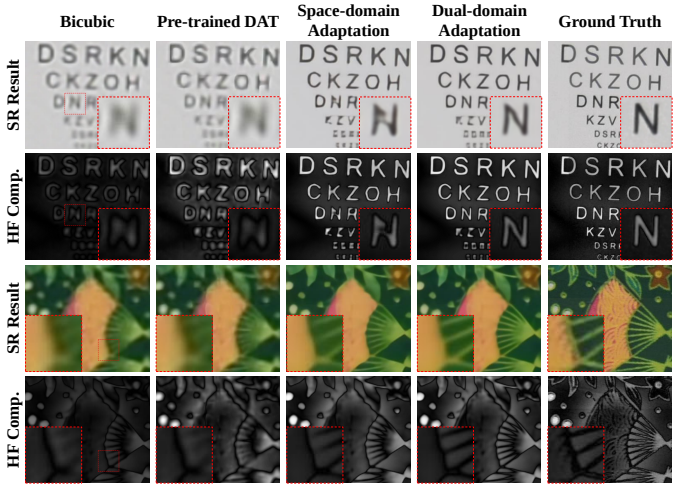

Fig. 1: Our dual-domain adaptation method aims to adapt pre-trained image super-resolution (SR) models from simulated to realistic datasets. The first and third rows show SR enhancements, and the second and fourth rows highlight improved high-frequency components. The second column presents the pre-trained DAT model outcomes [1]. The third and fourth columns demonstrate visual and structural improvements achieved with our spatial and dual-domain adaptation strategies, respectively.

I. Introduction

REALISTIC image super-resolution (SR) aims to transform low-resolution (LR) inputs from the real world to high-resolution (HR) images. In contrast to traditional super-resolution [2]–[4], which often deals with clean and synthetic datasets, realistic image SR focuses on handling real-world degradation patterns, making it much more applicable to actual applications. It focuses on enhancing overall image quality, recovering finer details and textures in a natural and perceptually appealing manner. This is important for improving user experiences in electronic devices such as TVs and smartphone photographs. It also has huge significance in applications such as security and surveillance, medical imaging, and forensic analysis, where improving image quality can lead to better decision-making, diagnosis, or analysis. Hence, this paper focuses on efficiently learning neural network-based models for addressing the realistic image SR task.

Recent methods [5], [6] are dedicated to improving the robustness of SR models by simulating the image degradation process to generate LR images from their HR counterparts. However, the real image degradation process remains more complex than these simulations. An emerging solution lies in harnessing directly captured LR-HR image pairs with diverse camera configurations [7]–[10]. The pioneering works [8]–[10] introduce methods focusing on leveraging adjacent information, improving the reconstruction of textural nuances, and modeling disparate high-frequency distributions. Considering image SR models pre-trained on large-scale simulated datasets are exposed to abundant images, they are advantageous at capturing these basic features. Intuitively, such prior knowledge of pre-trained image SR models can be used for improving the generalization ability, accelerating the training process, and relieving the burden of training data collection quality when coping with the realistic image SR task.

To address the challenges of realistic image SR, we

Manuscript received September 10, 2024. (*Corresponding author: De Cheng*)

Chaowei Fang and Bolin Fu are with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University, Xi'an, China. (e-mail: chaoweifang@outlook.com, 1179502349@qq.com)

De Cheng is with the School of Telecommunications Engineering, Xidian University, Xi'an, Shaanxi 710071, China (e-mail: dcheng@xidian.edu.cn).

Lechao Cheng is with School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China.

Guanbin Li is with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China (e-mail: liguanbin@mail.sysu.edu.cn).

propose a novel ***Dual-domain Adaptation Network*** (DAN) that efficiently adapts pre-trained SR models on simulated datasets to realistic scenarios. Our approach introduces two key innovations: a **spatial-domain adaptation strategy** and a **frequency-domain adaptation branch**, designed to synergistically enhance SR performance. Image SR models pre-trained on simulated datasets excel in extracting fundamental features, but their direct application to realistic datasets is suboptimal due to domain-specific variations. To bridge this gap, our spatial-domain adaptation strategy refines the pre-trained model by selectively freezing intermediate module parameters while updating others. This selective freezing strategy achieves an optimal balance between performance and retraining efficiency, outperforming the traditional approach of freezing only shallow modules. Additionally, we incorporate low-rank adaptation (LoRA) [11] to fine-tune static parameters with minimal computational overhead. These techniques enable effecctive and efficient adaptation of pre-trained SR models to realistic datasets. Fig. 1 provides two examples for illustrating the effect of our proposed method in adapting a pre-trained image SR model, DAT [1]. As shown by the third column of Fig. 1, the above spatial-domain adaptation strategy can significantly improve the SR results of the pre-trained DAT, while minimal extra trainable parameters are introduced.

High-frequency detail restoration is critical for high-quality SR but is inadequately addressed by spatial-domain techniques alone. This may lead to memorizing pixel values without explicitly capturing intricate high-frequency textural features. Previous researches [12]–[14] underscore the potential for image restoration based on the frequency domain, which exhibits advantages in recovering intricate high-frequency details compared to techniques based solely on the spatial domain. Exploiting this insight, we incorporate a frequency-domain adaptation branch into our model. Unlike existing frequency domain methodologies [13], [14], our approach merges the Fourier transform coefficients of the input images and intermediate features produced by the spatial-domain backbone model to infer HR frequency domain images. Subsequently, transforming these images to the spatial domain yields the final prediction. As illustrated by the fourth column of Fig. 1, the above frequency-domain adaptation branch helps enhance the SR results by generating more accurate and clear high-frequency components. Comprehensive evaluations of public realistic image SR benchmarks, including RealSR [8], D2CRealSR [10], and DRealSR [9], confirm the superior performances of our method over existing state-of-the-art realistic image SR solutions.

Our main contributions are summarized as follows:

- We introduce a dual-domain adaptation networks that is able to seamlessly transfer models from simulated to real-world image SR datasets.
- We devise a frequency-domain adaptation branch which can be flexibly integrated with existing spatial-domain backbone models, enhancing the restoration of high-frequency components.
- Through extensive benchmark testing on public realistic image SR datasets, our method establishes a new state-of-the-art in SR performances.

II. Related Work

A. Realistic Image Super-resolution

Image super-resolution, which seeks to increase the spatial resolution of images, attracts significant attention. Dong et al. [2] pioneer the introduction of the first deep neural networks (DNN) for the image SR task. With advances in DNN architectures, piles of models [1], [4], [15]–[30] are proposed with the aim of improving the performance of image SR.

To enhance the reality of image SR results, Ledig et al. [31] advocate for regularizing the VGG [32] feature distance and incorporating the generative adversarial learning loss [33] for constraining the optimization of network parameters. Fuoli et al. [12] present a Fourier space regularization loss to highlight the recovery of omitted frequency components. In particular, Liang et al. [34] dynamically detect visual anomalies in super-resolution images, emphasizing model learning in these regions. Li et al. [35] focus on addressing the conflict between perceptual and pixel-reconstruction-based objectives with exclusionary masks and devise a data distillation strategy to select simulated training data having similar noise patterns with the target dataset. Liu et al. [14] introduce a hybrid framework, interweaving spatial and frequency learning, employing spectral prediction uncertainty to combine the strengths of PSNR-oriented and adversarial learning-based SR models. However, a significant portion of existing SR research is heavily based on simulated training datasets, which are typically derived from image degradation operations such as blurring and interpolation. These operations oversimplify the intricate process inherent to real LR image formation, resulting in models that can not perform well on realistic datasets.

To address the challenge of realistic image SR, Chen et al. [7] collect real LR-HR image pairs by using various camera lenses to capture indoor postcard images. Both Cai et al. [8] and Wei et al. [9] expand on this, establishing more extensive realistic SR datasets spanning indoor and outdoor scenes. While these datasets focus on upsampling factors of 2, 3, and 4, Li et al. [10] introduce a $8\times$ SR dataset. The intricacies of capturing real-world images and ensuring pixel-level alignment make the creation of such datasets a laborious endeavor. Conversely, simulated image SR datasets are more straightforward to generate, and models pre-trained on them hold valuable insights that could improve realistic SR models. The efficient model adaptation from simulated to realistic datasets remains a relatively under-explored territory in image SR. This paper introduces a dual-domain adaptation network to bridge this gap. A spatial-domain adaptation strategy based on selective parameter updation and low-rank parameter adjustment is devised for efficiently transferring the pre-trained model. Moreover, a frequency-domain adaptation branch is incorporated for further amplifying the model's capability in high-frequency detail recovery.

B. Frequency-domain Representations for Image SR

Several recent methods focus on enhancing frequency-domain representations for image SR. Fuoli et al. [12] improve high-frequency content using Fourier-space supervision but primarily focus on perceptual quality with efficient models.

In contrast, our method combines spatial-domain adaptation and frequency-domain adaptation to refine both spatial and frequency features, preserving low-level details from a pre-trained model while enhancing high-frequency recovery. Wang et al. [13] explore the complementary nature of spatial and frequency domains in a two-branch network, but our approach goes further by adapting pre-trained models with selective parameter freezing in the spatial domain, enabling better generalization on realistic datasets. Liu et al. [14] focus on uncertainty estimation in the frequency domain, whereas our method directly improves high-frequency component recovery through Fourier domain adaptation while maintaining spatial domain performance. In summary, while existing methods focus on frequency-domain enhancements for SR, our approach integrates dual-domain adaptation to refine both spatial and frequency features. This not only improves high-frequency detail restoration but also preserves the low-level feature extraction capabilities of the pre-trained model, providing a more comprehensive solution for realistic image SR.

C. Neural Network Adaptation

Addressing domain-specific challenges, such as data imbalance [36], [37], few/noisy training data [38]–[40], and domain gap reduction [41]–[44], is widely studied in the field of image processing and understanding. The paradigm of transferring models pre-trained for generic tasks to specialized downstream tasks is an effective and economic solution to tackling this problem [45], [46]. A prevalent strategy involves fine-tuning the entire pre-trained model on the specific downstream tasks to achieve desirable results. To enhance efficiency and alleviate computational overhead during the fine-tuning phase, Yosinski et al. [47] and He et al. [48] advocate for the fine-tuning of only the terminal layers, while keeping preceding layers intact. Another line of research introduces adaptation modules directly into the neural network architecture. For instance, Houlsby et al. [49] and Chen et al. [50] recommend the inclusion of lightweight adaptation heads within intermediate layers. Chen et al. [51] optimize the model structure by pruning redundant structures and fusing multi-order graphs. Hu et al. [11] use a pair of down-projection and up-projection layers to adjust parameters of neural network layers without introducing computational overhead during inference. Zhang et al. [52] and Xu et al. [53] suggest the incorporation of supplementary side adaptation branches alongside original backbone architectures. A distinct approach, as showcased by Lester et al. [54] and Jia et al. [55], modifies pre-trained models for downstream tasks by injecting auxiliary information into the input signals. Recent studies have also tackled visual degradation and representation across diverse scenarios. For example, Huang et al. [56] address cross-modal retrieval via a two-stage asymmetric hashing method. Cheng et al. [57] propose a continual learning framework for all-in-one weather removal using knowledge replay, while Cheng et al. [58] introduce a contrastive learning strategy with progressive negative enhancement for robust image dehazing. Together, these works highlight the importance of adaptable models for handling real-world visual variations.

Inspired by these neural network adaptation techniques, we delve into transferring image SR models from simulated to realistic datasets with minimal effort. A dual-domain adaptation framework based on efficient spatial-domain parameter adjustment and frequency-domain feature integration is devised to implement the transfer of image SR models.

III. Methodology

This study addresses the challenge of realistic image SR. Let the input image be denoted as $\mathbf{X} \in \mathbb{R}^{c\times h\times w}$, where c, h, and w signify the number of channels, height, and width, respectively. A DNN-based image SR model is trained to predict a HR image, $\mathbf{O} \in \mathbb{R}^{c\times \rho h\times \rho w}$, from $\mathbf{X}$, with ρ being the upsampling ratio. The ground-truth HR image is represented as $\mathbf{Y}$.

A. Overview

To overcome the limitations of existing realistic image SR approaches [8]–[10] that struggle with generalizing basic features and reconstructing high-frequency details, we propose a novel framework called Dual-domain Adaptation Networks (DAN). By leveraging pre-trained SR models, our method integrates both spatial-domain adaptation (SDA) and frequency-domain adaptation (FDA) to improve generalization, accelerate training, and reduce reliance on extensive paired datasets. The framework is designed to efficiently adapt pre-trained models to real-world scenarios. An overview of our framework is shown in Fig. 2.

We use SwinIR [59] as the pre-trained backbone model to illustrate our approach, which consists of two key components:

1. **Spatial-Domain Adaptation (SDA):** This branch selectively fine-tunes specific layers of the pre-trained SR model while employing low-rank adaptation (LoRa) to efficiently adjust unselected layers. This ensures effective refinement without excessive parameter updates.
2. **Frequency-Domain Adaptation (FDA):** This branch focuses on restoring high-frequency components that are often lost in real-world low-resolution images, complementing the spatial-domain adaptations by explicitly addressing the high-frequency restoration challenges.

By combining these two strategies, the proposed framework preserves the pre-trained model's ability to capture fundamental features like edges and textures while enhancing its performance on complex, real-world degradations. This dual-domain approach delivers high-quality SR results with significantly reduced trainable network parameters.

B. Pre-trained Backbone Model

Models pre-trained on simulated datasets capture valuable prior knowledge for extracting basic image features necessary for super-resolving LR images. Leveraging this prior knowledge during training on realistic datasets accelerates the process, mitigates overfitting especially with limited realistic samples, and reduces computational costs by allowing certain parameters to remain fixed.

For illustration, we use SwinIR [59] as our backbone model. A convolutional layer first computes a preliminary feature

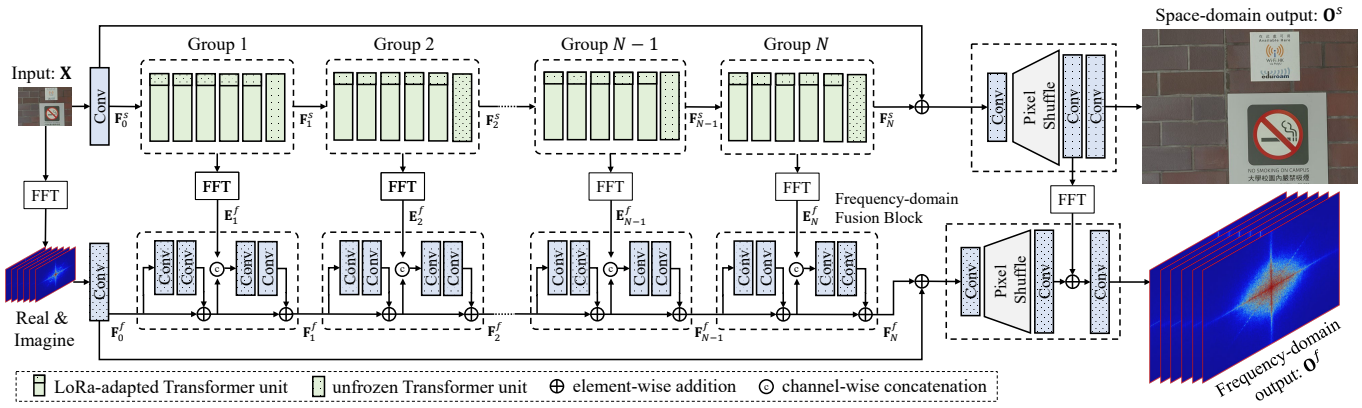

Fig. 2: Overview of our proposed dual-domain adaptation networks. It is built upon a pre-trained image SR model like SwinIR [59], which is constituted by a head convolution, N Transformer-based feature enhancement modules, and an upsampler. As shown in the upper stream, a spatial-domain adaptation strategy is introduced by unfreezing tail units of each feature enhancement module and applying low-rank adapters to adjust the remain units. The bottom stream presents the frequency-domain branch which progressively accumulates the spectral signals for enhancing the recovery of high-frequency components.

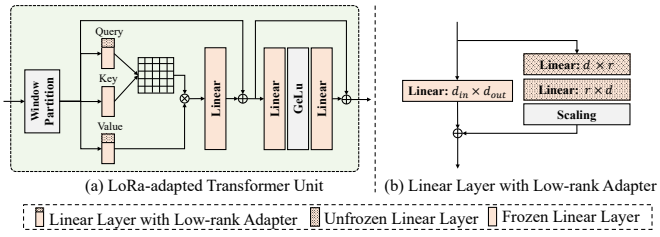

Fig. 3: (a) Within the adapted Transformer unit, low-rank adapters are incorporated to modify the parameters of the linear layers for generating the query and value variables. The workflow of the linear layer with low-rank adapter is illustrated in (b). The output of frozen vanilla linear layer in the left branch is adapted with residuals generated by a pair of down-projection and up-projection linear layers together with a scaling layer in the right branch.

map, $\mathbf{F}_0^s \in \mathbb{R}^{d\times h\times w}$, from the input $\mathbf{X}$, where d is the number of feature channels. This map is enhanced by N groups of Transformer units, each containing M units. We denote the intermediate feature maps produced by the n-th group of Transformer units as $\mathbf{F}_n^s \in \mathbb{R}^{d\times h\times w}$. Each Transformer unit partitions the input into local blocks. Then, it applies linear layers to compute query, key, and value variables, which are subsequently used to compute cross-pixel correlations to aggregate context information for feature enhancement. The final enhanced feature map is computed as $\mathbf{F}_{N+1}^s = \mathbf{F}_0^s + \mathbf{F}_N^s$, and an upsampler predicts the HR output $\mathbf{O}^s$ in the spatial domain.

While effective for simulated datasets, such models may underperform on realistic datasets due to differences in degradation patterns. To bridge this gap, we propose a dual-domain adaptation network (Fig. 2) that efficiently adapts the pre-trained model to realistic datasets with both spatial-domain and frequency-domain adaptation techniques.

C. Dual-domain Adaptation Networks

1) Spatial-Domain Adaptation.: Pre-trained models often encode general, low-level features such as edges, textures, and basic patterns in their initial layers. These foundational features are transferable across tasks and domains, making them essential for maintaining performance consistency. Freezing these layers during fine-tuning preserves this critical knowledge, preventing it from being overwritten and ensuring stable performance. Guided by this understanding, we freeze the parameters of the convolutional head and the first $M^{sta} \in \{0,1,2,\cdots,M\}$ units within each Transformer group. Conversely, the parameters of the remaining M^{dyn} units, where $M^{dyn} = M - M^{sta}$, are updated during training. This selective parameter freezing strategy enables the model to retain its ability to extract low-level features while adapting to task-specific nuances. Unlike completely freezing entire Transformer groups, which may introduce bottlenecks by constraining the propagation of domain-adaptive features, our devised strategy allows dynamic units to effectively process and integrate new information. This strategy mitigates overfitting and enhances the model's adaptability, ensuring improved performance on realistic datasets.

Inspired by Hu et al. [11], we integrate low-rank adapters to enhance the adaptability of frozen Transformer units. As delineated in Fig. 3, a pair of down-projection and up-projection linear layers are attached to each layer used for calculating query and value variables through the additive operation. We define r as the rank value of the adapter, with r being substantially smaller than d. By integrating these adapters, we built a robust spatial domain backbone model for addressing the realistic image SR task. Importantly, this configuration enables updates to only a restricted subset of network parameters and does not introduce additional computation demands during the inference phase.

2) Frequency-Domain Adaptation Branch.: The accurate restoration of high-frequency components is pivotal in the SR image reconstruction paradigm. We also observe that realistic

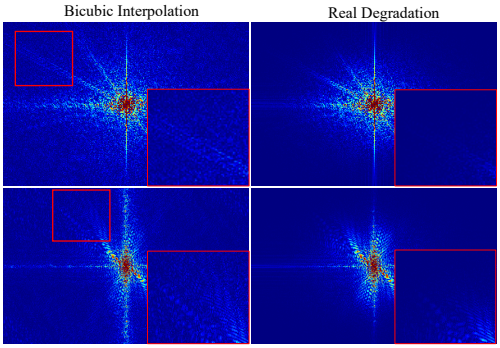

Fig. 4: Visualization of frequency signals of LR images simulated by bicubic interpolation (left) and realistic LR images (right).

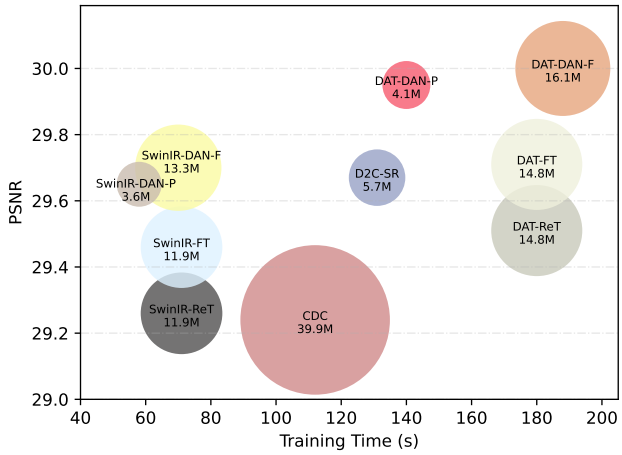

Fig. 5: Scatter plots of PSNR values, training time per epoch, and trainable parameter amount of different methods. Larger points indicate more trainable parameters.

LR images often lack the high-frequency signals found in those LR images simulated via bicubic interpolation. This difference is illustrated in Fig. 4, where each row shows the frequency amplitude maps of an LR image bicubicly interpolated from a HR image in the RealSR dataset [8] (left) and the corresponding real LR image (right). To enhance the capacity of the adapted SR model in restoring high-frequency components, we incorporate an auxiliary adaptation branch in the frequency domain, as depicted in the lower section of Fig. 2.

Let the real and imaginary components of the frequency domain for $\mathbf{X}$ be represented by $\mathbf{R}$ and $\mathbf{I}$, respectively. These components can be calculated utilizing the Fast Fourier Transformation (FFT) algorithm. A convolutional layer is used to extract the initial frequency domain feature $\mathbf{F}_0^f \in \mathbb{R}^{d^f \times h \times w}$ from the concatenation of $\mathbf{R}$ and $\mathbf{I}$.

Subsequently, we employ N fusion blocks to cumulatively combine the spectral data inherent in the intermediate features $\{\mathbf{F}_n^s\}_{n=1}^N$ of the spatial-domain backbone model. By FFT, we generate the spectral data $\mathbf{E}_n^f$, which is the concatenation of the real and imaginary components of $\mathbf{F}_n^s$. Each fusion block comprises two residual components: the initial component refines the frequency-domain feature map from the previous stage, namely $\mathbf{F}_{n-1}^f$, while the latter merges $\mathbf{E}_n^f$ into the refined feature map. This can be mathematically represented as:

$$\tilde{\mathbf{F}}_n^f = \mathbf{F}_{n-1}^f + \mathscr{F}_{res}(\mathbf{F}_{n-1}^f), \quad \mathbf{F}_n^f = \tilde{\mathbf{F}}_n^f + \mathscr{F}_{res}([\tilde{\mathbf{F}}_n^f, \mathbf{E}_n^f]), \tag{1}$$

where $\mathscr{F}_{res}(\cdot)$ signifies the forward function of the residual pathway, comprised of dual convolution layers.

As illustrated in the lower right quadrant of Fig. 2, an upsampling mechanism is adopted to increase the resolution of the frequency domain feature map. The feature map produced by the penultimate convolution layer of the space-domain backbone undergoes an FFT transformation and is subsequently fused into the frequency domain's upsampler using the additive operation. The resulting high-resolution spectral maps are denoted as $\mathbf{O}^f$, which yield the terminal super-resolution output $\mathbf{O}$ upon transformation into the spatial domain.

The combination between spatial-domain and frequency-domain adaptation plays a crucial role in enhancing the performance of image SR models. The spatial-domain adaptation (SDA) contributes to transferring the low-level feature extraction capabilities of the pre-trained model into realistic scenarios with the selective parameter adjustment strategy and low-rank adapters. Meanwhile, frequency-domain adaptation (FDA) directly addresses the loss of high-frequency components, which are essential for reconstructing fine details in realistic images. By combining these two strategies, SDA focuses on refining the spatial-domain representations of basic features, while FDA enhances the restoration of high-frequency information that may be ignored during training on realistic datasets.

D. Objective Function

The cost function utilized for parameter refinement comprises two principal facets: (i) the deviation between the output of the spatial domain backbone model, i.e., $\mathbf{O}^s$, and the ground-truth HR image $\mathbf{Y}$; and (ii) the discrepancy between the output of the frequency domain adaptation branch, i.e., $\mathbf{O}^f$, and the Fourier coefficients of $\mathbf{Y}$.

Both of these deviation terms are quantified using the L1 norm. Consequently, the composite cost function can be articulated as:

$$L = ||\mathbf{O}^s - \mathbf{Y}||_1 + \lambda ||\mathbf{O}^f - \mathrm{FFT}(\mathbf{Y})||_1, \tag{2}$$

where λ $(= 10)$ is a constant. Throughout the training phase, the modifiable parameters within the spatial-domain backbone model, along with entire parameters within the frequency-domain adaptation branch, undergo optimization.

IV. Experiments

A. Datasets

Three realistic SR datasets are used to evaluate the performance of SR methods:

- **RealSR** [8] contains LR and HR image pairs captured from 559 scenes using two DSLR cameras including Canon 5D3 and Nikon D810. Every scene has an HR

TABLE I: Comparison with existing methods on RealSR, D2CRealSR, and DRealSR datasets.

Method	RealSR						D2CRealSR		DRealSR					
	4×		3×		2×		8×		4×		3×		2×	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	27.24	0.764	28.61	0.810	31.67	0.887	27.74	0.822	30.56	0.822	31.50	0.835	32.67	0.877
SRResNet [31]	28.99	0.825	30.65	0.862	33.17	0.918	30.01	0.864	31.63	0.847	31.16	0.859	33.56	0.900
EDSR [18]	29.09	0.827	30.86	0.867	33.88	0.920	30.23	0.868	32.03	0.855	32.93	0.876	34.24	0.908
RCAN [19]	29.21	0.824	30.90	0.864	33.83	0.923	30.26	0.868	31.85	0.857	33.03	0.876	34.34	0.908
LP-KPN [8]	29.05	0.834	30.60	0.865	33.49	0.917	-	-	31.58	-	32.64	-	33.88	-
ESRGAN [60]	29.15	0.826	30.72	0.866	33.80	0.922	30.06	0.865	31.92	0.857	32.39	0.873	33.89	0.906
CDC [9]	29.24	0.827	30.99	0.869	33.96	0.925	30.02	0.841	32.42	0.861	33.06	0.876	34.45	0.910
D2C-SR [10]	29.67	0.830	31.28	0.870	34.30	0.925	30.47	0.869	31.79	0.852	32.69	0.865	34.07	0.904
SwinIR-PreT [59]	27.64	0.780	28.98	0.821	32.08	0.895	29.14	0.854	30.61	0.821	31.61	0.837	32.82	0.880
SwinIR-ReT	29.26	0.829	30.83	0.865	34.06	0.926	30.16	0.873	31.86	0.850	32.95	0.869	34.28	0.906
SwinIR-FT	29.46	0.833	31.14	0.871	34.16	0.926	30.24	0.875	32.23	0.856	33.10	0.872	34.55	0.909
SwinIR-DAN-P (Ours)	29.65	0.835	31.20	0.872	34.30	0.927	30.30	0.876	32.25	0.855	33.18	0.874	34.54	0.910
SwinIR-DAN-F (Ours)	29.70	0.837	31.31	0.874	34.40	0.928	30.36	0.876	32.45	0.859	33.34	0.877	34.75	0.913
DAT-PreT [1]	27.64	0.780	28.98	0.820	32.08	0.895	-	-	30.61	0.821	31.60	0.837	32.82	0.880
DAT-ReT	29.51	0.831	31.10	0.869	34.19	0.926	30.04	0.873	31.90	0.850	32.91	0.868	34.28	0.907
DAT-FT	29.71	0.839	31.36	0.876	34.41	0.929	30.26	0.877	32.32	0.860	33.18	0.874	34.62	0.912
DAT-DAN-P (Ours)	29.95	0.841	31.59	0.878	34.58	0.930	30.49	0.876	32.36	0.857	33.20	0.874	34.63	0.912
DAT-DAN-F (Ours)	**30.00**	**0.843**	**31.68**	**0.880**	**34.73**	**0.932**	**30.51**	**0.878**	**32.58**	**0.862**	**33.42**	**0.880**	**34.87**	**0.915**

image and its three LR counterparts having 1/2, 1/3, and 1/4 resolutions, respectively. 459 and 100 image pairs are used for training and testing, respectively.

- **D2CRealSR** [10] contains 115 pairs of LR and HR images for 8× image SR. They are split into 100 and 15 for training and testing, respectively.
- **DRealSR** [9] contains 884, 783, and 840 training image pairs, and 83, 84, and 93 testing image pairs for 2×, 3×, and 4× SR, respectively.

B. Implementation Details

In this study, we employ PyTorch [61] for the implementation of our proposed methodology. Our experimental framework involves the integration of dual-domain adaptation networks with two distinct backbone models: SwinIR [59] pre-trained on the DIV2K dataset, and DAT [1] pre-trained on a composite dataset combining DIV2K and Flickr2K. These backbone models are architecturally composed of six feature enhancement groups (i.e., $N = 6$), with each group comprising six Transformer units (i.e., $M = 6$). The intermediate feature maps within these models have 180 channels, i.e., $d = 180$. In our default configuration, we opt to freeze the parameters of the convolution head and the first five units of each Transformer group, namely $M^{sta} = 5$. Regarding the low-rank adapters, we assign a value of 4 to r. For DAT, low-rank adapters are applied to adjust the parameters of linear layers for generating key and value variables in spatial or channel-wise self-attention modules and those for constructing the spatial-gate feed-forward network. For the frequency-domain adaptation branch, the dimension d^f is set to 64. During training, we use the Adam optimizer [62] to update the parameters. The initial learning rate is established at 2×10^{-4}, which is subsequently halved every 2,000 iterations. Training images are processed to randomly crop 96×96 LR patches. We set the batch size at 4 and conduct the training for 70,000 iterations.

C. Comparisons with Existing Methods

1) Quantitative Comparison: In our comprehensive evaluation, detailed in Table I, we perform a comparative analysis using PSNR and SSIM metrics to assess our dual-domain adaptation methodology against established image SR techniques such as SRResNet [31], EDSR [18], RCAN [19], LP-KPN [8], ESRGAN [60], CDC [9], D2C-SR [10], along with variants SwinIR or DAT adapted from simulated datasets to realistic ones. The metric values of SRResNet and EDSR are taken from [10]. This assessment highlights that pre-trained models including SwinIR-PreT and DAT-PreT, show reduced effectiveness on realistic datasets. Attempts to re-train SwinIR and DAT from scratch as indicated by SwinIR-ReT and DAT-ReT, respectively, exhibit a notable underperformance against D2C-SR across various SR settings on the RealSR dataset.

We try to explore full fine-tuning (FT) of all parameters on the pre-trained SwinIR and DAT, forming SwinIR-FT and DAT-FT, respectively. These two models consistently improve performance across all SR settings. Despite its effectiveness,

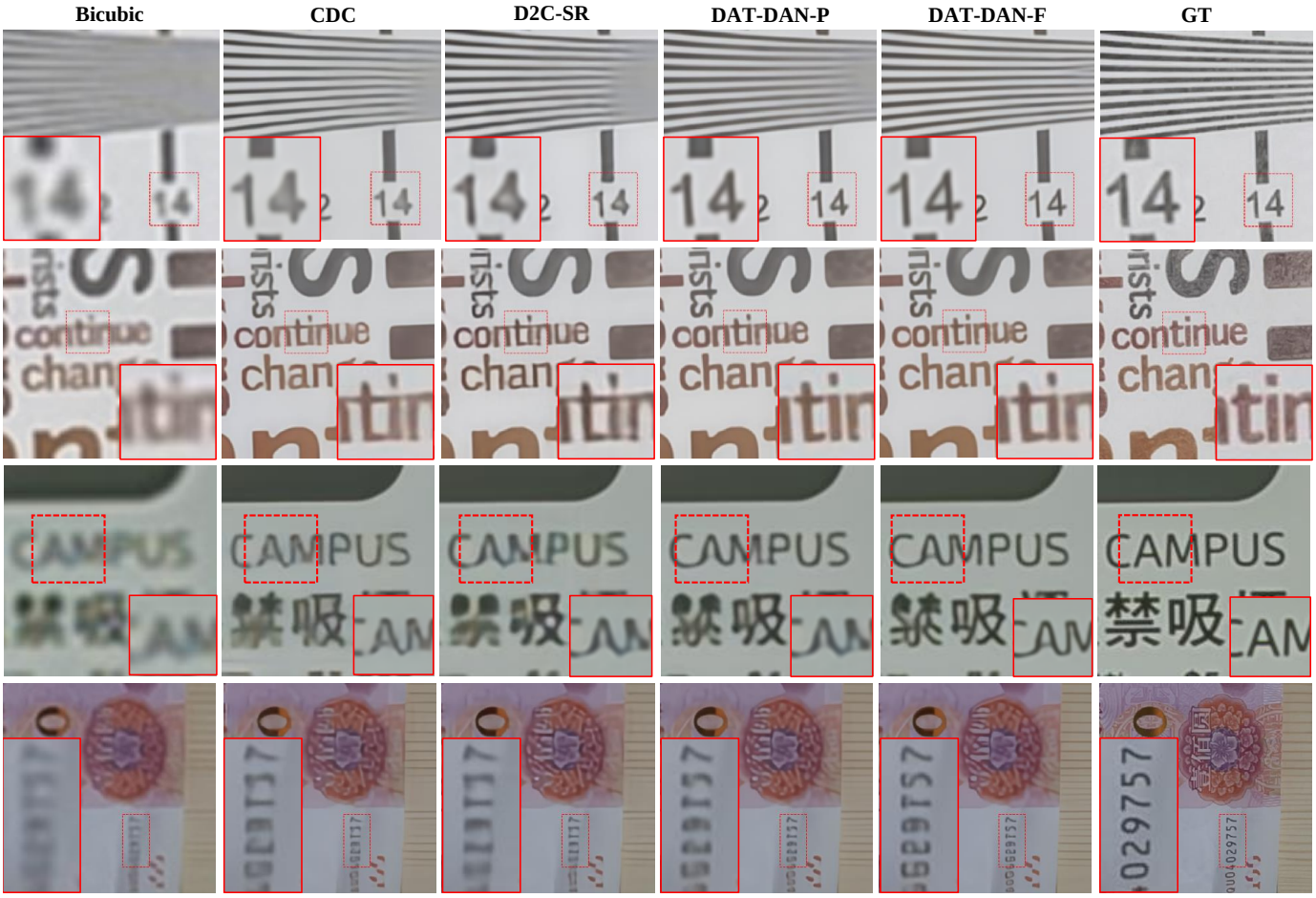

Fig. 6: A qualitative comparison among different models on four 4× SR examples from the RealSR dataset. From left to right: bicubic interpolation, CDC [9], D2C-SR [10], DAT-DAN-P, DAT-DAN-F, and the ground-truth (GT) image.

the FT process incurs computational inefficiencies due to the necessity of updating all parameters. We then integrate our novel dual-domain adaptation networks with the pre-trained SwinIR and DAT, leading to the development of SwinIR-DAN-P and DAT-DAN-P, respectively. These variants, which require around 30% of the trainable parameters of their FT counterparts, respectively, demonstrate not only a reduction in the number of trainable parameters but also outperforming the FT method, with SwinIR-DAN-P and DAT-DAN-P surpassing SwinIR-FT and DAT-FT by 0.21dB and 0.24dB, respectively, in the 4× SR setting. In particular, DAT-DAN-P exceeds the performance of the previously best method, D2C-SR. Our method's potential is further underscored by unfreezing all parameters in the backbone models, resulting in SwinIR-DAN-F and DAT-DAN-F variants, which show substantially improved results.

Fig. 5 shows the scatter plot of PSNR values, training time per epoch, and trainable parameter amount of different methods. All metrics are evaluated under the 4× SR setting of the RealSR dataset. The training time per epoch is tested with a Nvidia GeForce RTX 3090 GPU. It can be observed that our method variant SwinIR-DAN-P or DAT-DAN-P have much less training time and trainable parameters while achieving higher PSNR value, compared to SwinIR-FT or DAT-FT, respectively. This indicates that our devised DAN is effective in improving the training efficiency compared to full fine-tuning. Compared to D2C-SR, SwinIR-DAN-P has significantly less training time while achieving comparable PSNR value; DAT-DAN-P has comparable training time while leading to much higher PSNR value. It can be deduced that making use of image SR models pre-trained with simulated data can effectively improve the SR performance or training efficiency for learning realistic image SR models.

2) Qualitative Comparison: Fig. 6 and Fig. 7 visualize examples from the different SR settings of RealSR, DRealSR and D2CRealSR dataset, respectively. As illustrated by the four examples of Fig. 6, our method variants DAT-DAN-P and DAT-DAN-F are capable of generating characters with sharper edges and clearer appearances than D2C-SR and CDC. As illustrated of Fig. 7, our method demonstrates enhanced performance in reconstructing building surface streaks and high-frequency structural details

D. Cross-camera Adaptation

To validate the robustness of our method in the situation of cross-camera adaptation, we conduct experiments by dividing the RealSR dataset into two subsets according to the camera, including Canon and Nikon. One subset is used to pre-train the DAT model which is subsequently adapted to the other one. The experimental results are presented in Table II. Our

Fig. 7: Qualitative comparison among different models on $2\times, 3\times, 4\times$ SR setting of the DRealSR dataset [9] and $8\times$ SR setting of the D2CRealSR [10]. From left to right: bicubic interpolation, CDC [9], D2C-SR [10], DAT-DAN-P, DAT-DAN-F, and the ground-truth (GT) image.

TABLE II: Performance in cross-camera testings on RealSR dataset, including Canon→Nikon and Nikon→Canon. DAT is used as the backbone model.

Settings	Metrics	ReT	FT	DAN-P
Canon→Nikon	PSNR	28.11	28.28	28.53
	SSIM	0.794	0.811	0.814
Nikon→Canon	PSNR	29.23	29.35	29.60
	SSIM	0.832	0.839	0.840

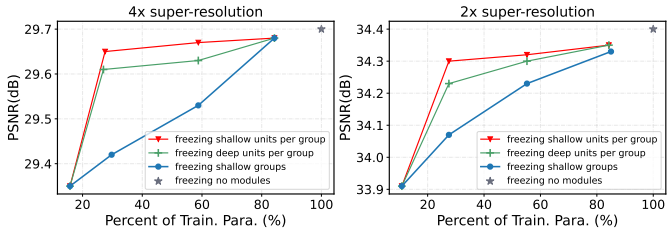

Fig. 8: Variation of PSNR with respect to the percent of trainable parameters.

proposed method DAN-P performs significantly better than retraining from scratch (ReT) and full fine-tuning (FT). For example, under the setting of Canon→Nikon, the PSNR of DAN-P is 0.42 and 0.25 higher than that of ReT and FT, respectively. These experiments further validate the generalization of our method in different real-world conditions, affirming the robustness of our DAN-P and DAN-F variants.

E. Ablation Study

This section is dedicated to conducting comprehensive ablation experiments to validate the effectiveness of the key components in our method. We also evaluate the impact of employing diverse strategies and hyper-parameters for model adaptation. These experiments are carried out systematically on the RealSR dataset, and SwinIR is used as the space-domain backbone model.

• **Efficacy of Principal Components.** Table III illustrates the implementation of various variants of our method to substantiate the effectiveness of key components including selective parameter fine-tuning (SPFT), low-rank adapters (LoRa), and frequency-domain adaptation (FDA). The pre-trained SwinIR model serves as the foundational baseline (referenced as No.

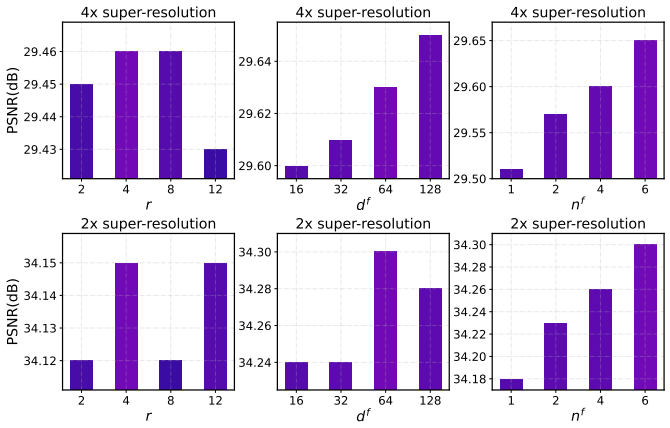

Fig. 9: Variation of the PSNR metric with respect to the rank value r, the feature dimension d^f in the frequency-domain adaptation branch and the number of frequency domain adaptation stage n^f.

TABLE III: Ablation study for key components in 4× and 2× settings of RealSR. 'SPFT', 'LoRa', and 'FDA' denotes selective parameter fine-tuning, low-rank adapter, and frequency-domain adaptation, respectively.

No.	SPFT	LoRa	FDA	4×		2×	
				PSNR	SSIM	PSNR	SSIM
1				29.26	0.829	34.06	0.926
2	✓			29.47	0.831	34.11	0.925
3	✓	✓		29.46	0.832	34.15	0.925
4	✓		✓	29.58	0.834	34.24	0.926
5	✓	✓	✓	29.65	0.835	34.30	0.927

1). Upon fine-tuning selective parameters of this baseline model, encompassing the parameters of the last Transformer units in all feature enhancement groups and the upsampler (referenced as No. 2), an enhancement in the SR results is observed. For example, there is an increase of 0.21dB in the PSNR metric in the 4× SR setting.

Furthermore, the incorporation of FDA, as illustrated in No. 4, results in substantial performance improvement. Specifically, there is an increase in the PSNR metric by 0.11dB and 0.13dB in the 4× and 2× SR settings, respectively. As depicted in No. 3 and No. 5, the use of LoRa to modulate the frozen parameters of the backbone model contributes to certain performance enhancements. We attempt to alter the FDA branch with an extra spatial-domain branch having more parameters. As shown in Table IV, this method variant indicated by (SDA-P) is unable to improve the PSNR and SSIM metrics compared to No. 3 in Table III, indicating the extra spatial-domain branch is redundant in adapting the backbone model. However, our devised FDA branch can still bring improvement by enhancing the recovery of high-frequency components in the Fourier domain.

TABLE IV: Performance of replacing the FDA branch in DAN-P with a spatial domain adaptation branch (SDA-P) on 4× RealSR dataset.

Variants	PSNR	SSIM	N^{trn}
SDA-P	29.44	0.832	4.6
DAN-P	29.65	0.835	3.6

TABLE V: The results of replacing FFT with wavelet transform on the RealSR dataset.

Method Variants	4×		2×	
	PSNR	SSIM	PSNR	SSIM
Wavelet Transform	29.47	0.832	34.15	0.925
Fast Fourier Transform	29.65	0.836	34.30	0.927

• **Comparison FFT against Wavelet Transform in Spatial-Frequency Decomposition.** We implement a variant of our method through replacing the Fast Fourier Transform (FFT) with wavelet transform (WT). The results of this variant are presented in the third row of Table V. Compared to FFT, WT shows a performance decline, e.g., the PSNR of WT is 0.18 and 0.15 lower than that of FFT on 4× and 2× SR settings, respectively. The potential reason is that FFT provides a global frequency decomposition which is useful for capturing periodic patterns, textures, and high-frequency details in a holistic manner while WT only captures local high-frequency details. This global perspective is particularly advantageous in SR tasks that require precise restoration of high-frequency components.

• **Choice of Parameter Freezing Strategies.** Our analysis extends to the performance implications of employing three distinct strategies for freezing parameters of the backbone model. In the first strategy, we freeze the parameters of the shallowest groups completely. We label this strategy as "freezing shallow groups" in Fig. 8. In the second strategy which is labeled as "freezing deep units per group" in Fig. 8, we freeze the parameters of deep Transformer units of each group while updating the parameters of other Transformer units. The third strategy, which is the one adopted in our final method, involves freezing the first several Transformer units in each group. As indicated by the curve labeled "freezing shallow units per group" in Fig. 8, this approach yields a more favorable balance between performance and the number of trainable parameters.

• **Variability of Hyper-parameters.** The influence of choosing values for the rank value r, the feature dimension d^f within the frequency-domain adaptation branch and the number of frequency-domain adaptation stages (denoted as n^f) is depicted in Fig. 9.

1) The rank value r controls the number of learnable parameters in adapters for frozen layers. Optimal SR performance is achieved by setting r to 4, while larger values for r do not confer additional benefits.
2) The parameter d^f determines the complexity of the FDA branch. The peak performance in 4× and 2× SR is attained at different values. In the 2× SR context, the

Fig. 10: PSNR variation curves produced with different numbers of training images.

PSNR value plateaus beyond a d^f setting of 64; in the $4\times$ SR context, the PSNR value shows a continuous increase as d^f increases from 16 to 128. Balancing both performance and resource consumption considerations, we establish d^f at 64 in the final version of our method.

3) Increasing the number of FDA stages n^f leads to higher PSNR values, since using more FDA stages helps capture more nuanced high-frequency features.

• **Performance under Various Numbers of Training Images.** We conduct experiments using various numbers of training images including 10, 25, 50, and 100. Fig. 10 illustrates the PSNR variation curves of FT and DAN-P with respect to the training iteration. The PSNR values of FT degrade substantially with increasing training iteration due to overfitting. The overfitting issue intensifies as the number of training images decreases. The PSNR values of our DAN-P remain stably as the training process advances when using 100 training images. For 50 training images, the PSNR value of DAN-P suffers a decrease at around 3,000 iterations and then saturates in subsequent training iterations. For 25 and 10 training images, moderate overfitting issue of DAN-P can be observed. In summary, compared to FT, our DAN-P method can effectively alleviate the overfitting issue while achieving better performance.

V. Conclusion

This study introduces a dual-domain adaptation network for transferring image SR models from simulated to realistic datasets. We find that selective parameter fine-tuning and frequency domain adaptation notably improve SR performance. Our analysis reveals that freezing intermediate Transformer units offers a better performance-resource balance compared to freezing the shallowest modules. Low-rank adapters also contribute to adjusting the frozen parameters of the backbone model. Our devised network adaptation method significantly outperforms the full fine-tuning strategy using nearly one third of the trainable parameters. Our method can significantly improve pre-trained backbone models, achieving new state-of-the-art performances on RealSR, D2CRealSR, and DRealSR datasets. The limitation of this work is that the pre-trained image SR models usually have high network complexity. Learning light-weight image SR models with the help of the knowledge of pre-trained large models deserves future research.

References

[1] Z. Chen, Y. Zhang, J. Gu, L. Kong, X. Yang, and F. Yu, "Dual aggregation transformer for image super-resolution," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 12 312–12 321.

[2] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, 2015.

[3] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 2472–2481.

[4] X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong, "Activating more pixels in image super-resolution transformer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 22 367–22 377.

[5] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, "Designing a practical degradation model for deep blind image super-resolution," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 4791–4800.

[6] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-esrgan: Training real-world blind super-resolution with pure synthetic data," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 1905–1914.

[7] C. Chen, Z. Xiong, X. Tian, Z.-J. Zha, and F. Wu, "Camera lens super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 1652–1660.

[8] J. Cai, H. Zeng, H. Yong, Z. Cao, and L. Zhang, "Toward real-world single image super-resolution: A new benchmark and a new model," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 3086–3095.

[9] P. Wei, Z. Xie, H. Lu, Z. Zhan, Q. Ye, W. Zuo, and L. Lin, "Component divide-and-conquer for real-world image super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 101–117.

[10] Y. Li, H. Huang, L. Jia, H. Fan, and S. Liu, "D2c-sr: A divergence to convergence approach for real-world image super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2022, pp. 379–394.

[11] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv:2106.09685*, 2021.

[12] D. Fuoli, L. Van Gool, and R. Timofte, "Fourier space losses for efficient perceptual image super-resolution," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 2360–2369.

[13] C. Wang, J. Jiang, Z. Zhong, and X. Liu, "Spatial-frequency mutual learning for face super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 22 356–22 366.

[14] T. Liu, J. Cheng, and S. Tan, "Spectral bayesian uncertainty for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 18 166–18 175.

[15] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 1874–1883.

[16] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 1637–1645.

[17] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 624–632.

[18] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2017, pp. 136–144.

[19] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 286–301.

[20] C. Fang, G. Li, X. Han, and Y. Yu, "Self-enhanced convolutional network for facial video hallucination," *IEEE Trans. Image Process.*, vol. 29, pp. 3078–3090, 2019.

[21] F. Zhu, C. Fang, and K.-K. Ma, "Pnen: Pyramid non-local enhanced networks," *IEEE Trans. Image Process.*, vol. 29, pp. 8831–8841, 2020.

[22] C. Ma, Y. Rao, J. Lu, and J. Zhou, "Structure-preserving image super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7898–7911, 2021.

[23] Y. Liu, S. Wang, J. Zhang, S. Wang, S. Ma, and W. Gao, "Iterative network for image super-resolution," *IEEE Trans. Multimedia*, vol. 24, pp. 2259–2272, 2021.

[24] C. Fang, D. Zhang, L. Wang, Y. Zhang, L. Cheng, and J. Han, "Cross-modality high-frequency transformer for mr image super-resolution," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 1584–1592.

[25] H. Choi, J. Lee, and J. Yang, "N-gram in swin transformers for efficient lightweight image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 2071–2081.

[26] W. Li, J. Li, G. Gao, W. Deng, J. Zhou, J. Yang, and G.-J. Qi, "Cross-receptive focused inference network for lightweight image super-resolution," *IEEE Trans. Multimedia*, vol. 26, pp. 864–877, 2023.

[27] X. Zhou, H. Huang, Z. Wang, and R. He, "Ristra: Recursive image super-resolution transformer with relativistic assessment," *IEEE Trans. Multimedia*, 2024.

[28] D. Liu, X. Wang, R. Han, N. Bai, J. Hou, and S. Pang, "Cte-net: Contextual texture enhancement network for image super-resolution," *IEEE Trans. Multimedia*, 2024.

[29] R. Ran, L.-J. Deng, T.-J. Zhang, J. Chang, X. Wu, and Q. Tian, "Knlconv: Kernel-space non-local convolution for hyperspectral image super-resolution," *IEEE Trans. Multimedia*, 2024.

[30] Q. Liu, P. Gao, K. Han, N. Liu, and W. Xiang, "Degradation-aware self-attention based transformer for blind image super-resolution," *IEEE Trans. Multimedia*, 2024.

[31] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 4681–4690.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, 2014.

[33] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Proc. Adv. Neural Inform. Process. Syst.*, vol. 27, 2014.

[34] J. Liang, H. Zeng, and L. Zhang, "Details or artifacts: A locally discriminative learning approach to realistic image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 5657–5666.

[35] H. Li, J. Qin, Z. Yang, P. Wei, J. Pan, L. Lin, and Y. Shi, "Real-world image super-resolution by exclusionary dual-learning," *IEEE Trans. Multimedia*, vol. 25, pp. 4752–4763, 2022.

[36] S. Sun, S. Zhi, Q. Liao, J. Heikkilä, and L. Liu, "Unbiased scene graph generation via two-stage causal modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12 562–12 580, 2023.

[37] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of big data*, vol. 6, no. 1, pp. 1–54, 2019.

[38] C. Wen, H. Huang, Y. Ma, F. Yuan, and H. Zhu, "Dual-guided frequency prototype network for few-shot semantic segmentation," *IEEE Trans. Multimedia*, 2024.

[39] Y. Song, T. Wang, P. Cai, S. K. Mondal, and J. P. Sahoo, "A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities," *ACM Comput. Surv.*, vol. 55, no. 13s, pp. 1–40, 2023.

[40] Y. Chen, Y. Wu, N. Han, X. Fang, B. Chen, and J. Wen, "Partial multi-label learning based on near-far neighborhood label enhancement and nonlinear guidance," in *Proc. ACM Int. Conf. Multimedia*, ser. MM '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 3722–3731. [Online]. Available: https://doi.org/10.1145/3664647.3681300

[41] X. Fang, N. Han, G. Zhou, S. Teng, Y. Xu, and S. Xie, "Dynamic double classifiers approximation for cross-domain recognition," *IEEE Trans. Cybern.*, vol. 52, no. 4, pp. 2618–2629, 2020.

[42] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018.

[43] S. Wang, Q. Zang, D. Zhao, C. Fang, D. Quan, Y. Wan, Y. Guo, and L. Jiao, "Select, purify, and exchange: A multisource unsupervised domain adaptation method for building extraction," *IEEE Trans. Neural Netw. Learn. Syst.*, 2023.

[44] G. Li, C. Fang, Z. Chen, M. Mao, and L. Lin, "Uncertainty-aware active domain adaptive salient object detection," *IEEE Trans. Image Process.*, 2024.

[45] Y. Guo, H. Shi, A. Kumar, K. Grauman, T. Rosing, and R. Feris, "Spottune: transfer learning through adaptive fine-tuning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 4805–4814.

[46] X. Li, Y. Grandvalet, F. Davoine, J. Cheng, Y. Cui, H. Zhang, S. Belongie, Y.-H. Tsai, and M.-H. Yang, "Transfer learning in computer vision tasks: Remember where you come from," *Image Vis. Comput.*, vol. 93, p. 103853, 2020.

[47] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *Proc. Adv. Neural Inform. Process. Syst.*, vol. 27, 2014.

[48] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 16 000–16 009.

[49] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2019, pp. 2790–2799.

[50] S. Chen, C. Ge, Z. Tong, J. Wang, Y. Song, J. Wang, and P. Luo, "Adaptformer: Adapting vision transformers for scalable visual recognition," *Proc. Adv. Neural Inform. Process. Syst.*, vol. 35, pp. 16 664–16 678, 2022.

[51] Y. Chen, R. Chen, Q. Li, X. Fang, J. Li, and W. K. Wong, "Denoising high-order graph clustering," in *Proc. IEEE Int. Conf. Data Eng.* IEEE, 2024, pp. 3111–3124.

[52] J. O. Zhang, A. Sax, A. Zamir, L. Guibas, and J. Malik, "Side-tuning: a baseline for network adaptation via additive side networks," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 698–714.

[53] M. Xu, Z. Zhang, F. Wei, H. Hu, and X. Bai, "Side adapter network for open-vocabulary semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 2945–2954.

[54] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," *arXiv:2104.08691*, 2021.

[55] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2022, pp. 709–727.

[56] J. Huang, P. Kang, N. Han, Y. Chen, X. Fang, H. Gao, and G. Zhou, "Two-stage asymmetric similarity preserving hashing for cross-modal retrieval," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 1, pp. 429–444, 2024.

[57] D. Cheng, Y. Ji, D. Gong, Y. Li, N. Wang, J. Han, and D. Zhang, "Continual all-in-one adverse weather removal with knowledge replay on a unified network structure," *IEEE Trans. Multimedia*, 2024.

[58] D. Cheng, Y. Li, D. Zhang, N. Wang, J. Sun, and X. Gao, "Progressive negative enhancing contrastive learning for image dehazing and beyond," *IEEE Trans. Multimedia*, 2024.

[59] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 1833–1844.

[60] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. Workshop*, 2018, pp. 0–0.

[61] S. Imambi, K. B. Prakash, and G. Kanagachidambaresan, "Pytorch," *Programming with TensorFlow: Solution for Edge Computing Applications*, pp. 87–104, 2021.

[62] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.