Learning Prompt Adapters for Forgetting-Free Continual Image Super-Resolution

Chaowei Fang, *Member, IEEE*, Bolin Fu, De Cheng, Chengpei Tang, and Guanbin Li, *Member, IEEE*

***Abstract*—Continual image super-resolution (CISR) aims to efficiently adapt a pre-trained model to a variety of tasks while retaining knowledge from previously learned tasks, minimizing the need for intensive independent training. The primary challenges include *catastrophic forgetting* due to varying data distributions and degradation types, along with the necessity for high adaptability. While prompt-based continual learning has proven effective in image classification, its direct application to super-resolution (SR) often fails to meet the demands for detailed pixel-level restoration and domain discrimination in low-level characteristics. To address these challenges, we propose Learning Prompt Adapters (LPA), which dynamically generates pixel-wise prompts through a combination of multi-granularity prompt bases and identities. By adaptively integrating these prompts into the Transformer architecture, we effectively improve the model's performance on fine-grained details in super-resolution tasks, as well as enhancing the model's adaptability to new tasks and preserving knowledge from previous ones. Through organizing the low-rank prompt bases with specific identities, we set up an effective solution to managing cross-task differences and enhancing prompt richness. Extensive experiments on benchmarks comprising the NYU, RealSR, DIV2K, REDS, and MANGA109 datasets with diverse degradation types demonstrate that LPA significantly outperforms existing continual learning methods. Codes of this paper are available at: https://github.com/dummerchen/LPA**

***Index Terms*—Continual learning, image super-resolution, neural network adaptation.**

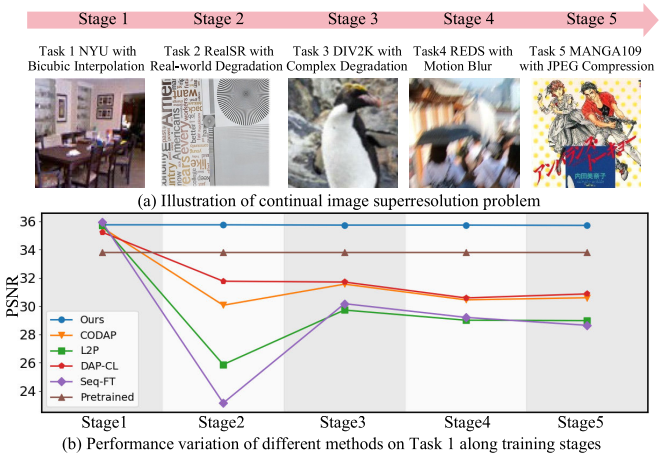

Fig. 1. (a) Illustrates the continual image SR challenge, where a pre-trained SR model is sequentially adapted to five tasks with distinct image domains and degradation processes. (b) shows PSNR metrics for the first task across training stages. The pre-trained model struggles to generalize, and sequential fine-tuning (Seq-FT) or existing continual learning methods (Dap-CL [1], L2P [2], CODA-P [3]) fail to prevent catastrophic forgetting. Our method, *Learning Prompt Adapters*, effectively mitigates this issue.

I. Introduction

IMAGE super-resolution (SR) has made substantial progress in fields like natural image processing. However, existing SR methods [4], [5] are typically task-specific and struggle to generalize across different domains and degradation types. In real-world scenarios, the types of images needing SR evolve over time, as exemplified by Fig. 1(a), due to varying image capture conditions. Continually retraining models for each new task is computationally expensive. Hence, we propose continual image super-resolution (CISR), a framework that allows pre-trained SR models to adapt efficiently to new tasks while retaining prior knowledge. Unlike task-specific methods, CISR ensures scalability, maintains consistent performance across evolving SR tasks, and reduces the need for constant retraining, making it particularly suited for dynamic environments where image domains and degradation types vary.

CISR introduces unique challenges beyond those seen in traditional continual learning tasks such as image classification or semantic segmentation, which primarily focus on high-level features. In contrast, SR requires the pixel-level restoration of fine-grained details, which vary significantly across tasks due to differences in scenes, imaging methods, and domains. Managing domain discrepancies—caused by variations in low-level image characteristics such as resolution, texture, and color profiles—complicates the transfer of knowledge between tasks. Moreover, SR tasks involve different degradation types, from simple bicubic downsampling to complex unknown distortions, which creates a larger variance in input space

Received 17 March 2025; revised 18 January 2026; accepted 23 February 2026. Date of publication 12 March 2026; date of current version 17 March 2026. This work was supported in part by the National Key Research and Development Program of China under Grant 2024YFB3908503; in part by the National Natural Science Foundation of China under Grant 62376206, Grant 62322608, and Grant 62576262; in part by the Key Research and Development Program of Shaanxi Province under Grant 2024GX-YBXM135; and in part by the Fundamental Research Funds for the Central Universities under Grant QTZX25083. The associate editor coordinating the review of this article and approving it for publication was Dr. Zhengxia Zou. *(Corresponding author: De Cheng.)*

Chaowei Fang and Bolin Fu are with the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education, School of Artificial Intelligence, Xidian University, Xi'an 710071, 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).

Chengpei Tang is with the School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen, Guangdong 518107, 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).

Digital Object Identifier 10.1109/TIP.2026.3671688

compared to classification tasks. As illustrated in Fig. 1(b), we attempt to sequentially fine-tune a pre-trained SR model (SwinIR [6]) on four SR tasks. The experimental images of four tasks are selected from NYU, RealSR, DIV2K, and REDS datasets, respectively. Their low-resolution images are generated through bicubic interpolation, real-world degradation, complex degradation, and motion blur, respectively. We find that the performance of the SR model on the earliest task degrades rapidly, underscoring the difficulty of retaining SR quality over time. Efforts of applying existing classification-oriented continual learning methods (Dap-CL [1], L2P [2], and CODA-P [3]) are unable to alleviate the catastrophic forgetting issue as well. This highlights the need for tailored continual learning strategies that address the specific challenges of CISR.

Prompt tuning [7], [8] is a kind of lightweight adaptation techniques that aim to efficiently adapt pre-trained models for specific downstream applications. Rather than retraining the entire model, prompt tuning modifies a small set of learnable prompt tokens, allowing the model to adapt to various tasks without significant computational overhead. In the field of image restoration, prompt tuning has been applied to adjust backbone models for different distortion scenarios, including varying levels of compression and diverse degradation types and intensities [9], [10], [11]. These works show that prompt tuning can efficiently adapt a pre-trained model to different restoration challenges by leveraging task-relevant prompts, making it a suitable approach for continual image super-resolution (CISR). However, CISR introduces three key challenges that prompt tuning must address. On one hand, to memorize the knowledge of historically learned image SR tasks, we need to restore the prompts sequentially learned for each task. On the other, since the task prior for a given test image is unknown, retrieving the most appropriate prompts is critical for performance. Additionally, SR tasks demand fine-grained feature adaptation at the pixel level, which requires a more nuanced application of prompt tuning compared to high-level tasks like classification.

To address these challenges, we introduce *Learning Prompt Adapters* (LPA), a novel method specifically designed to enhance task adaptability and knowledge retention. Unlike conventional prompt-based continual learning methods [1], [2], [3] which insert learnable prompt tokens ahead to tune the backbone model, LPA dynamically generates prompt adapters to tune the internal layers for each input image of the image SR backbone model, through combining multi-granularity prompt bases guided by specific identities. Our prompt adapters directly modulate the key and value projections within the basic self-attention modules in the SR model, enabling pixel-level and feature-level adaptation in intermediate layers. This design is particularly suitable for low-level visual knowledge preservation, as it allows fine-grained adjustments to the feature representations that are critical for pixel-wise restoration in SR tasks. The adapter-based prompt design enables the creation of pixel-wise prompt information. This capability significantly improves the model's ability to restore fine-grained details when learning new tasks, making LPA particularly effective for super-resolution applications.

During the continuous learning process, these prompt bases and identities are incrementally learned and restored for each image SR task. In particular, the prompt identities are formed by clustering the semantic and degradation-aware features of each task's training images. The semantic features are extracted with the image encoder of CLIP [12] which is learned via image-text alignment, while the degradation identity features are constructed based on a dedicated degradation encoder, which is specifically designed to capture low-level degradation attributes. This dual-identity design enables the model to differentiate task-specific domains or degradations and retrieve the corresponding prompt bases. It provides an effective solution to managing task differences caused by varying image domains or degradation types. Moreover, to achieve the accurate retrieval of the accumulated prompt bases, we construct the prompt adapters using the dual feature similarities between each input image and prompt identities to aggregate those prompt bases. This ensures that the appropriate prompts are dynamically assigned to input images, thereby enhancing adaptability in diverse scenarios. Extensive testing on benchmarks utilizing the NYU [13], RealSR [5], DIV2K [14], REDS [15], and MANGA109 [16] datasets demonstrates LPA's superiority against existing continual learning methods.

Our main contributions are summarized as below.

- We present a novel method, *Learning Prompt Adapters* (LPA), that dynamically generates pixel-wise prompts, enhancing the model's adaptation ability to restore fine-grained details in CISR.
- Our approach utilizes multi-granularity prompt bases and identities to manage domain shifts caused by varying visual content and degradation processes, ensuring appropriate prompt assignment for diverse input images.
- Extensive experiments on three benchmarks demonstrate that LPA outperforms existing methods.

II. Related Work

A. Image Super-Resolution

Recent advancements in image SR focus on restoring low-resolution (LR) images based on specific image data domain and degradation pattern. We will briefly introduce different approaches for addressing these image SR tasks. For the image SR task in which LR images are synthesized with bicubic interpolation, early approaches in image super-resolution adopt various deep convolutional neural network (CNN) architectures [4], [17], which achieve excellent performance. However, CNN-based architectures face challenges in capturing global dependencies across spatial and channel dimensions effectively. Reference [6] improve the network architectures of image SR through exploring non-local context information with shifted window Transformers [18]. ELAN [19] further improves the self-attention mechanism by using multiple window sizes to capture long-distance correlations. Reference [20] relies on the permutation operation for compressing spatial information into channel information, thus enlarging the windows size without bringing much computation burden. Other methodologies such as [21] and [22] endeavor to integrate spatial and channel-wise self-attention

mechanism for exploring both local and global correlations, effectively improving the image SR performance.

Various approaches are developed to address image SR tasks where LR images are synthesized using bicubic interpolation. Initially, deep convolutional neural networks (CNNs) are employed, achieving notable results [4], [6], [19], [21], [22], [23], [24], [25], [26].

Practically, LR images often result from complex degradation types mixing blurring, downsampling, and noise injection. Various studies like [27], [28], [29], and [30] develop frameworks to handle multiple degradation types, guided by specific blur kernels. Some researchers, including [5] and [31], construct training pipelines and real-world datasets to better simulate and capture the actual degradation type, thus improving model generalization to realistic conditions. Considering LR images often suffer from motion blurring, a few works [32], [33] devise algorithms to address the image deblurring and super-resolving jointly. Despite these advancements, existing SR models often fail to adapt to shifts in image domains or degradation types. This paper proposes an efficient approach to adapt an image SR model to multiple tasks sequentially, enabling it to tackle new tasks without forgetting previous capabilities.

B. Continual Learning

Continual learning focuses on training a single model across different tasks while addressing catastrophic forgetting. Regularization-based approaches [34], [35], [36] are a classical type of methods proposed for continual learning. However, these methods can compromise adaptability in image SR due to shared backbone models. Other strategies include network expansion [37], [38], which increases capacity for new tasks, and rehearsal-based methods [39], [40] that retain a subset of old task data. Prompt-based methods have emerged as effective in managing task transitions. Techniques such as learning-to-prompt [2], which use a pool of prompts to adapt the backbone model, and dual prompt systems [3], [41] that differentiate between task-invariant and task-specific knowledge, help balance knowledge preservation and adaptability. InsVP [42] generates both pixel-space instance prompts and token-level prompts from the input image, thereby enhancing the modeling of instance-specific characteristics. TCPA [43] shows that effective prompt learning in Vision Transformers (ViT) requires fine-grained attention coordination between prompt tokens and image tokens.

Recent work [44] introduced SEC-Prompt for continual learning, which relies on class supervision and multi-class-per-task settings, with prompt losses defined by inter-class similarity. CL-LoRA [45] extends LoRA [46] to continual learning by maintaining multiple task-specific low-rank adapters and explicitly selecting the corresponding adapter at test time based on task identity. MoE-PromptCL [47] reinterprets ViT attention as an implicit Mixture-of-Experts (MoE), where prefix prompt tuning corresponds to introducing task-specific experts for new tasks, and further incorporates non-linear residual gating into prompt-related attention branches to facilitate more effective expert selection. However, these approaches are primarily designed for classification tasks and rely on discrete prompt tokens or prefix prompts to modulate high-level representations. While effective for semantic adaptation, these prompts operate at a coarse granularity and are not tailored for pixel-level restoration.

We introduce a prompt-based approach specifically designed for continual image SR challenges. Our method innovatively uses multi-granularity prompt bases and identities to capture task prompt information and generates adapters dynamically from the prompt bases for tuning the backbone model.

C. Domain Adaptation for Image Super-Resolution

To advance the application of image SR models in real-world scenarios, some methods employ unsupervised domain adaptation to transfer SR models from source to target datasets. Approaches like [48] and [49] use adversarial learning to reduce the domain gap. Although both domain adaptation and continual learning techniques enhance performance on new tasks, they address different challenges. Domain adaptation aims to reduce domain shifts using both source and target data to align feature spaces, while continual learning handles sequential tasks without prior data, managing current task data only.

III. Methodology

Suppose we have M image SR training tasks, denoted as $\{\mathbb{T}_1, \mathbb{T}_2, \cdots, \mathbb{T}_M\}$. Each task consists of a dataset with pairs of low-resolution (LR) images and their high-resolution (HR) counterparts. $\mathbb{T}_m$ is represented as $\{(\mathbf{X}_i^{(m)}, \mathbf{Y}_i^{(m)})\}_{i=1}^{N_m}$, where N_m is the number of image pairs, and $\mathbf{X}_i^{(m)}$ and $\mathbf{Y}_i^{(m)}$ are the LR and HR images, respectively. The goal is to learn single SR model on these tasks sequentially, minimizing resource and computation costs. During testing, the model should super-resolve input images without needing the task identity.

A. Overview

To tackle the CISR task, we propose a prompt-tuning method called *Learning Prompt Adapters* that enables a universal SR model to perform well across multiple SR tasks. Our method focuses on generating dynamic pixel-wise prompts to adapt the image SR backbone effectively.

As shown in Fig. 2, the framework comprises three main components: an image SR backbone, multi-granularity task bases and identities, and prompt adapters. We use the pre-trained SwinIR [6] as the backbone model, tuning its internal blocks with prompt adapters tailored to each task. These adapters generate pixel-wise prompt information that enhances the backbone model's intermediate features. The prompt adapters are built using a series of prompt bases, organized with multi-granularity task identities to handle cross-task domain variations. These identities, derived from features of the CLIP image encoder [12] and the degradation encoder [50], ensure diverse and effective prompt retrieval for each task. Each prompt adapter is dynamically generated by combining these prompt bases in a weighted manner for each image.

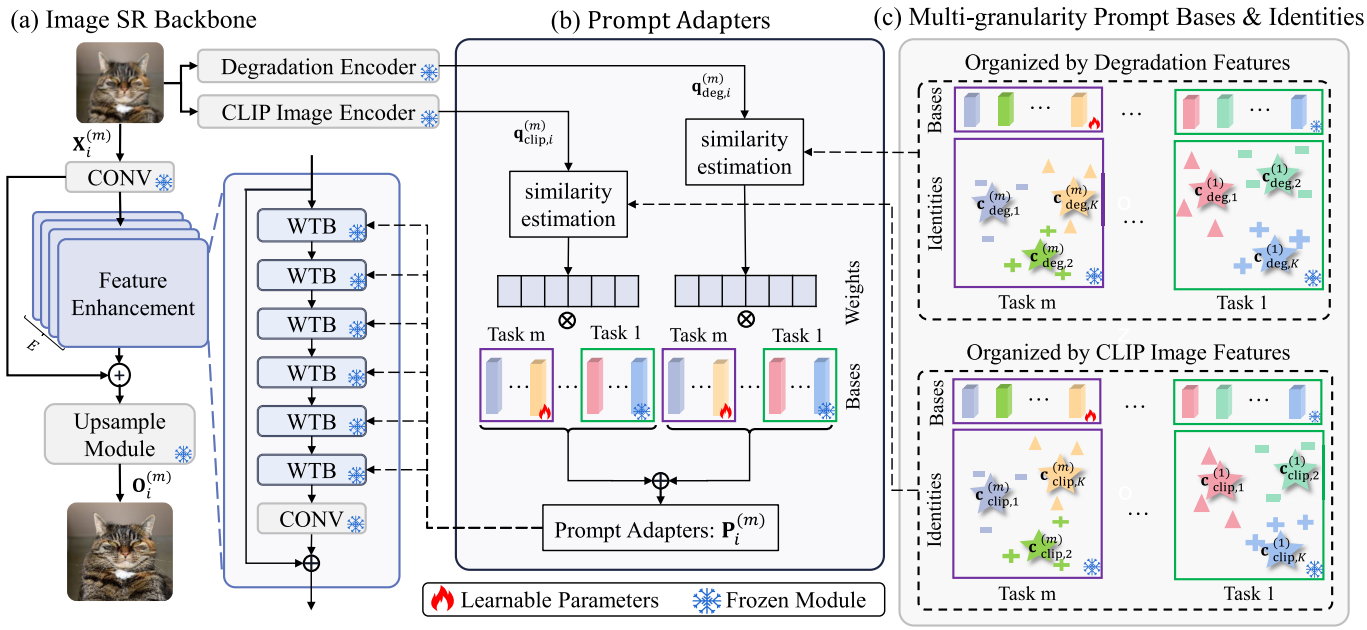

Fig. 2. Overview of our framework for adapting the pre-trained image SR backbone model with prompt adapters. The architecture of the image SR backbone model is shown in (a), with the prompt adapters applied for task adaptation. (b) The prompt adapters are constructed through a weighted combination of multi-granularity prompt bases. These prompt bases capturing task-specific prompt information are restored and signified with identities constructed with semantic and degradation feature centroids as shown in (c).

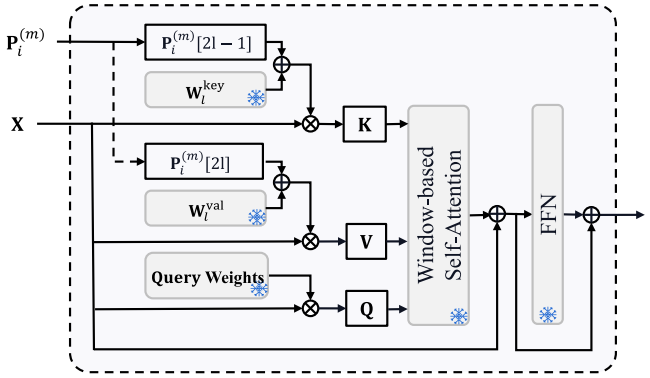

Fig. 3. The prompt adapters are used to adjust the parameters of key and value layers in the transformer module, achieving the generation of pi xel-wise prompt information.

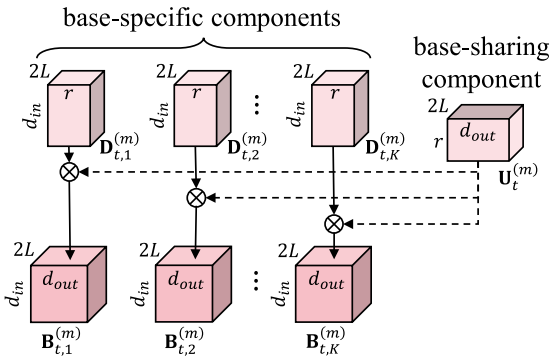

Fig. 4. The prompt bases are formed by multiplying a series of base-specific components with a base-sharing component.

B. Image SR Backbone

Our approach adapts the SwinIR backbone, which includes a head convolution layer, a feature enhancement module with a long-distance skip connection, and an upsample module, as illustrated in Fig. 2 (a). The feature enhancement module consists of E stages, each with six window-based transformer blocks, named 'WTB'. In each Transformer, query, key, and value variables are computed with linear layers for implementing the self-attention mechanism.

Starting from a SwinIR backbone pre-trained on a source SR task, we adapt it to new tasks by dynamically generating pixel-wise prompts for each input image. This is achieved by using adapters to adjust the key and value layers within Transformers. For an input image $\mathbf{X}_i^{(m)}$, we generate prompt adapters $\mathbf{P}_i^{(m)} \in \mathbb{R}^{2L \times d_{\text{in}} \times d_{\text{out}}}$. L represents the number of Transformers to be adapted. d_{in} and d_{out} represent input and output dimensions of linear layers, respectively. Let the original weights for key and value linear layers of the l-th module be $\mathbf{W}_l^{\text{key}}$ and $\mathbf{W}_l^{\text{val}}$, respectively. We adjust these weights by adding the respective prompt terms, forming adapted weights $\mathbf{W}_{m,i,l}^{\text{key}}$ and $\mathbf{W}_{m,i,l}^{\text{val}}$:

$$\mathbf{W}_{m,i,l}^{\text{key}} = \mathbf{W}_l^{\text{key}} + \mathbf{P}_i^{(m)}[2l-1], \tag{1}$$

$$\mathbf{W}_{m,i,l}^{\text{val}} = \mathbf{W}_l^{\text{val}} + \mathbf{P}_i^{(m)}[2l]. \tag{2}$$

Given an input feature $\mathbf{F}$, the adapted key and value layers produce $\mathbf{K}$ and $\mathbf{V}$, respectively:

$$\mathbf{K} = \mathbf{F}\mathbf{W}_{m,i,l}^{\text{key}} = \mathbf{F}\mathbf{W}_l^{\text{key}} + \mathbf{F}\mathbf{P}_i^{(m)}[2l-1], \tag{3}$$

$$\mathbf{V} = \mathbf{F}\mathbf{W}_{m,i,l}^{\text{val}} = \mathbf{F}\mathbf{W}_l^{\text{val}} + \mathbf{F}\mathbf{P}_i^{(m)}[2l]. \tag{4}$$

The first terms represent the original key and value, while the second terms add task-specific prompt information, adapting the backbone model to varying SR tasks through pixel-wise prompts. Unlike methods [2], [3] relying on prefix prompts, this direct modulation of key and value projections within the Transformer's window attention mechanism enables pixel-level feature adaptation in intermediate layers, while maintaining the ability to restore fine-grained details.

C. Learning Prompt Adapters

The core of our method, *Learning Prompt Adapters*, allows the SR backbone to adapt to different SR tasks by meeting two goals: generating task-specific prompts for each input image dynamically and maintaining fine-grained prompt diversity. These objectives are achieved by constructing the prompt adapters through a combination of multi-granularity prompt bases, guided by specific identities.

1) Multi-Granularity Prompt Bases and Identities: For each SR task, we create multiple prompt bases to store prompt information. These prompt bases are organized with two identity types: one derived from CLIP image features to capture semantic characteristics, and another based on degradation features to capture low-level image qualities. Each prompt base is represented by a matrix $\mathbf{B}_{t,k}^{(m)} \in \mathbb{R}^{2L\times d_{\text{in}}\times d_{\text{out}}}$, where t indicates the identity type ('clip' for semantic features or 'deg' for degradation features), and k indexes the bases within each task. The total number of bases for each task is defined as K.

To minimize trainable parameters, we factorize each prompt base matrix $\mathbf{B}_{t,k}^{(m)}$ into two low-rank matrices, $\mathbf{D}_{t,k}^{(m)} \in \mathbb{R}^{2L\times d_{\text{in}}\times r}$ and $\mathbf{U}_{t}^{(m)} \in \mathbb{R}^{2L\times r\times d_{\text{out}}}$, so that $\mathbf{B}_{t,k}^{(m)} = \mathbf{D}_{t,k}^{(m)}\mathbf{U}_{t}^{(m)}$. $\mathbf{D}_{t,k}^{(m)}$ is specific to each prompt base, encoding fine-grained knowledge within the task, while $\mathbf{U}_{t}^{(m)}$ is shared across bases, capturing common knowledge at the task level. This decomposition reduces the parameters needed per task while preserving the diversity of prompt information stored.

We also need an effective way to retrieve the most relevant prompt bases, given the domain differences across SR tasks. We achieve this by associating each prompt base with a specific task identity which is constructed by clustering features extracted exclusively from the training images of each task. CLIP image features [12] are used to form identities based on high-level semantics. Degradation features [50] are constructed based on a dedicated degradation encoder that is specifically designed to capture low-level degradation attributes. These identities enable the model to differentiate task-specific domains and degradations and retrieve the corresponding prompt bases. Importantly, both types of features are extracted only from the training set, and no test images or test-time features are involved in the clustering process. The clustering process is implemented using the K-means algorithm [51], a widely adopted and effective clustering method. The feature centroids of these clusters serve as prompt identities $\mathbf{c}_{t,k}^{(m)}$, with $\mathbf{c}_{\text{clip},k}^{(m)}$ and $\mathbf{c}_{\text{deg},k}^{(m)}$ respectively representing semantic and degradation identities for the prompt bases.

These multi-granularity prompt identities facilitate the retrieval of relevant prompt information by capturing the semantic and degradation feature distributions of each task. The prompt identities formed by distinct feature centroids can guide the prompt bases to learn fine-grained prompt information for each task, allowing the prompt adapters to generate fine-grained adjustments for the image SR model.

2) Prompt Adapter Construction: To construct prompt adapters, we aggregate prompt bases using weights derived from the similarity between the input image and prompt identities. For an input image $\mathbf{X}_i^{(m)}$, similarities to prompt identities are calculated based on features from the CLIP image encoder and degradation encoder.

We first calculate query features from $\mathbf{X}_i^{(m)}$ using the respective encoders to obtain query feature. Specifically, let $\mathbf{q}_{t,i}^{(m)}$ denote the query feature extracted from $\mathbf{X}_i^{(m)}$ by the CLIP image encoder or the degradation encoder. This query feature vector is used to compare with the identity centroids $\mathbf{c}_{t,k}^{(j)}$ for retrieving relevant prompt bases. The similarity for the k-th prompt of the j-th task with CLIP-based identity (t = 'clip') is computed using cosine distance:

$$\text{sim}_{\text{clip}}(\mathbf{q}_{\text{clip},i}^{(m)}, \mathbf{c}_{\text{clip},k}^{(j)}) = \frac{e^{\cos(\mathbf{q}_{\text{clip},i}^{(m)}, \mathbf{c}_{\text{clip},k}^{(j)})}}{\sum_{j'=1}^{m}\sum_{k'=1}^{K} e^{\cos(\mathbf{q}_{\text{clip},i}^{(m)}, \mathbf{c}_{\text{clip},k'}^{(j')})}}. \tag{5}$$

For degradation-based identity (t = 'deg'), Euclidean distance is used to estimate the similarity:

$$\text{sim}_{\text{deg}}(\mathbf{q}_{\text{deg},i}^{(m)}, \mathbf{c}_{\text{deg},k}^{(j)}) = \frac{e^{\|\mathbf{q}_{\text{deg},i}^{(m)} - \mathbf{c}_{\text{deg},k}^{(j)}\|_2^{-2}}}{\sum_{j'=1}^{m}\sum_{k'=1}^{K} e^{\|\mathbf{q}_{\text{deg},i}^{(m)} - \mathbf{c}_{\text{deg},k'}^{(j')}\|_2^{-2}}}. \tag{6}$$

The final prompt adapter $\mathbf{P}_i^{(m)}$ is constructed by aggregating the prompt bases using the similarity scores as weights:

$$\mathbf{P}_i^{(m)} = \sum_{t\in\{\text{'clip'},\text{'deg'}\}} \sum_{j=1}^{m}\sum_{k=1}^{K} \text{sim}_t(\mathbf{q}_{t,i}^{(m)}, \mathbf{c}_{t,k}^{(j)})\mathbf{B}_{t,k}^{(j)}. \tag{7}$$

This formulation ensures that the prompt adapter emphasizes the most relevant information for the current task. By using both the CLIP image features and degradation features, the model effectively captures semantic and low-level degradation differences between tasks.

The design allows the model to retrieve task-specific prompt information while also leveraging previously learned prompt bases. This not only prevents catastrophic forgetting but also promotes the reuse of shareable knowledge across tasks. Consequently, the prompt adapter can adapt flexibly to new tasks while maintaining fine-grained restoration capabilities, enhancing both generalization and detailed recovery in CISR.

D. Training and Inference Procedures

During training for the m-th task, we first generate prompt identities centroids by clustering features extracted from task's training images using the CLIP image encoder and degradation encoder. We then learn task-specific prompt bases by minimizing the L_1 loss:

$$L = \frac{1}{N_m}\sum_{i=1}^{N_m} \|\mathbf{O}_i^{(m)} - \mathbf{Y}_i^{(m)}\|_1, \tag{8}$$

where $\mathbf{O}_i^{(m)}$ is the SR output for $\mathbf{X}_i^{(m)}$, and $\mathbf{Y}_i^{(m)}$ is the ground truth. During training, only the current task's prompt bases

TABLE I

PERFORMANCE ON THREE CISR BENCHMARKS. AVERAGED PSNR, SSIM, AND FPSNR ON ALL TASKS ARE USED FOR EVALUATION. SEQ-FT IS SHORT FOR SEQUENTIAL FINE-TUNING

Scale	Methods	N-b → R-r → D-c → E-m → M-j diff. datasets under diff. degrad.			R-b → R-c → R-r → R-j same dataset under diff. degrad.			N-c → R-c → D-c →E-c → M-c diff. datasets under same degrad.		
		PSNR↑	SSIM↑	FRMSE↓	PSNR↑	SSIM↑	FRMSE↓	PSNR↑	SSIM↑	FRMSE↓
x2	Individual-Training	36.09	0.920	-	38.57	0.939	-	28.82	0.812	-
	Joint-Training	35.37	0.913	-	38.40	0.939	-	29.43	0.816	-
	Seq-FT	33.83	0.889	0.4157	36.63	0.893	0.3451	27.83	0.781	0.1401
	BSRGAN (Pretrained)	30.32	0.871	-	30.89	0.888	-	27.81	0.776	-
	SwinIR (Pretrained)	32.74	0.877	-	36.11	0.884	-	25.47	0.648	-
	LoRa [46]	33.92	0.888	0.4023	36.65	0.893	0.3007	27.78	0.787	0.1031
	L2P [2]	33.61	0.888	0.4126	36.58	0.893	0.2798	27.77	0.788	0.0836
	DualP [41]	33.56	0.887	0.4120	36.56	0.893	0.2747	27.61	0.785	0.0810
	CODA-P [3]	33.76	0.887	0.3650	36.43	0.892	0.2555	27.60	0.777	0.0774
	Dap-CL [1]	33.57	0.885	0.2971	36.33	0.890	0.2027	27.47	0.766	0.0743
	Ours	**34.53**	**0.905**	**0.0419**	**36.78**	**0.919**	**0.1340**	**27.83**	**0.789**	**0.0700**
x4	Individual-Training	29.57	0.831	-	31.46	0.858	-	25.14	0.718	-
	Joint-Training	29.12	0.819	-	31.28	0.858	-	25.59	0.723	-
	Seq-FT	27.30	0.789	0.4571	29.67	0.814	0.3141	24.38	0.701	0.1033
	BSRGAN (Pretrained)	26.14	0.758	-	27.51	0.785	-	24.43	0.679	-
	SwinIR (Pretrained)	26.18	0.746	-	28.67	0.794	-	23.26	0.605	-
	LoRa [46]	27.43	0.792	0.3566	29.66	0.815	0.2727	24.32	0.702	0.0914
	L2P [2]	27.20	0.787	0.3938	29.58	0.814	0.2582	24.37	0.702	0.0809
	DualP [41]	27.24	0.788	0.3600	29.63	0.815	0.2480	24.36	0.700	0.0708
	CODA-P [3]	27.35	0.787	0.2754	29.48	0.813	0.2105	24.36	0.695	0.0565
	Dap-CL [1]	27.23	0.781	0.2101	29.39	0.815	0.1513	24.33	0.687	0.0533
	Ours	**28.57**	**0.805**	**0.0098**	**30.02**	**0.834**	**0.1006**	**24.44**	**0.703**	**0.0490**

are updated while prompt bases learned from previous tasks remain frozen to prevent forgetting. After training task m, we save the learned prompt bases along with the corresponding identity centroids for later retrieval. During testing, given an input image, we extract its query features, compute similarities to the stored identity centroids, and dynamically aggregate prompt bases from all learned tasks using Eq. (7) to construct the prompt adapters. These adapters are then applied to adjust the key and value linear layers of Transformers as in Eqs. (1) and (2).

IV. EXPERIMENTS

A. Datasets and Benchmarks

To simulate real-world variety, we conduct experiments on four datasets: NYU, RealSR, DIV2K, REDS, and MANGA109, covering indoor, outdoor, and cartoon environments. NYU [13] comprises 1449 indoor images, split into 795 for training and 654 for testing. RealSR [5] consists of 559 LR and HR image pairs, with 400 pairs for training and 100 for testing. DIV2K [14] comprises 800 high-resolution training images and 100 test images. The REalistic and Dynamic Scenes (REDS) dataset [15] is widely used for video deblurring and super-resolution tasks, comprising 240 training sequences, 30 validation sequences, and 30 test sequences, totaling 30,000 video frames. MANGA109 [16] consists of

TABLE II

ABLATION STUDY OF PRINCIPLE COMPONENTS FOR CONSTRUCTING PROMPT ADAPTERS IN THE FIRST CISR SETTING

No.	CLIP	Degradation	Single	Multi	PSNR	SSIM	FRMSE
1					27.30	0.789	0.4571
2	✓		✓		27.54	0.778	0.0025
3	✓			✓	27.61	0.780	0.0024
4		✓	✓		28.15	0.792	0.0169
5		✓		✓	28.23	0.798	0.0104
6	✓	✓	✓		28.35	0.797	0.0172
7	✓	✓		✓	**28.57**	**0.805**	**0.0098**

TABLE III

PERFORMANCE OF DIFFERENT DESIGNS TO CONSTRUCT PROMPT BASES ON THE FIRST CISR BENCHMARK

Method Variants	PSNR	SSIM
separate $\mathbf{D}$, separate $\mathbf{U}$	28.00	0.795
separate $\mathbf{D}$, shared $\mathbf{U}$	**28.57**	**0.805**

TABLE IV

ABLATION STUDY ON DISTANCE METRICS FOR SIMILARITY ESTIMATION. HERE, **Cosine** DENOTES COSINE SIMILARITY, AND **Euclidean** DENOTES EUCLIDEAN DISTANCE. RESULTS ARE REPORTED ON THE FIRST CISR BENCHMARK

Similarity Metrics		Average		Task1 N-b		Task2 R-r		Task3 D-c		Task4 E-m		Task5 M-j	
CLIP ID	Degradation ID	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Euclidean	Euclidean	27.58	0.791	31.00	0.917	27.56	0.776	24.29	0.610	27.94	0.786	27.15	0.864
Cosine	Cosine	27.53	0.792	30.44	0.914	27.53	0.777	24.29	0.611	**28.07**	**0.789**	**27.32**	**0.867**
Cosine	Euclidean	**28.57**	**0.805**	**35.72**	**0.949**	**28.51**	**0.806**	**24.84**	**0.648**	27.75	0.783	26.05	0.837

TABLE V

PERFORMANCE OF METHOD VARIANTS INTEGRATED WITH DIFFERENT BACKBONES

Backbone	Method	Task1 N-b		Task2 R-r		Task3 D-c		Task4 E-m		Task5 M-j		Average	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DAT [21]	Seq-FT	28.80	0.903	27.43	0.772	24.22	0.606	**28.05**	**0.791**	**28.30**	**0.879**	27.36	0.790
	Ours	**34.29**	**0.944**	**27.76**	**0.783**	**24.56**	**0.624**	27.84	0.784	26.04	0.845	**28.10**	**0.796**
HAT [54]	Seq-FT	28.89	0.901	27.39	0.772	24.21	0.605	**28.02**	**0.790**	**28.89**	**0.890**	27.48	0.792
	Ours	**34.11**	**0.942**	**28.57**	**0.807**	**24.81**	**0.648**	27.89	0.788	26.33	0.842	**28.34**	**0.805**
SwinIR [6]	Seq-FT	28.65	0.899	27.41	0.772	24.22	0.603	**28.06**	**0.792**	**28.14**	**0.881**	27.30	0.789
	Ours	**35.72**	**0.949**	**28.51**	**0.806**	**24.84**	**0.648**	27.75	0.783	26.05	0.837	**28.57**	**0.805**

109 manga images, with the first 100 images allocated for training and the remaining 9 images for testing.

Our evaluation benchmarks involve five commonly-used degradation methods to generate LR images for different image super-resolution tasks, including: 1) bicubic interpolation (b); 2) real-world degradation (r) which is the default setting of the RealSR dataset; 3) complex degradation formed by random combinations as in [31] (c); 4) motion blurring, video compressing, and downscaling (m); 5) JPEG compressing and downscaling (j). We denote '[X]-[Y]' as an image SR task in which the degradation type Y is applied to generate LR images for the dataset X. Here, $X \in$ {'NYU' ('N'), 'RealSR' ('R'), 'DIV2K' ('D'), 'REDS' ('E'), 'Manga109' ('M')}, and $Y \in$ {'b', 'r', 'c', 'm', 'j'}. We set up three CISR benchmarks with SR tasks formed by varying datasets and degradation types: **1) Different datasets under different degradation**: This benchmark comprises five tasks including N-b, R-r, D-c, E-m, and M-j in order, which are different in both datasets and degradation types. **2) Same dataset under different degradation**: This benchmark comprises four tasks including R-b, R-c, R-r, and R-j in order, which share the same dataset and have different degradation types. **3) Different datasets under same degradation**: This benchmark comprises five tasks including N-c, R-c, D-c, E-c, and M-c in order, which have different datasets and share the same degradation type.

The above benchmarks reflect the challenges of image domain and degradation diversity, providing a comprehensive framework for assessing CISR methods' adaptability and performance. For each benchmark, we evaluate different methods on 2× and 4× SR tasks.

For quantitative evaluation, we calculate PSNR and SSIM on the Y channel of the YCbCr space. We calculate the forgetting ratio from the RMSE values denoted as **FRMSE** as follows. FRMSE is a task-agnostic scalar metric that summarizes forgetting behavior in super-resolution. It is inspired by the widely used average forgetting measure in continual learning [52], but is computed using RMSE to better reflect restoration fidelity. The calculation expression of FRMSE is as follows:

$$\text{FRMSE} = \frac{1}{M-1}\sum_{m=1}^{M-1}\frac{\text{RMSE}_m^{(M)} - \text{RMSE}_m^{(m)}}{\text{RMSE}_m^{(m)}}, \tag{9}$$

where $\text{RMSE}_m^{(m')}$ represents RMSE of each method on the m-th task after the m'-th training stage.

B. Implementation Details

We implement our method based on PyTorch, using Adam [53] optimizer with a decay rate of 0.9 and a learning rate of 4×10^{-4}. For training, we randomly crop all input images to 32×32 patches, and the model is trained for 6.4×10^4 iterations with a batch size of 32. By default, d_{in} and d_{out} are equal to 180. E, K and r is set to 6, 3 and 12, respectively. For the inference phase, we use a model that is continuously learned across all tasks. To facilitate efficient processing, the input image is divided into 32×32 patches through cropping and processed independently by the network. The final output is reconstructed by aggregating all processed patches.

C. Comparison With Existing Methods

To demonstrate the upper bound for the three CISR benchmarks, we first implement Individual-Training, which learns a specific model for each task, and evaluate a unified model trained on all tasks' data, named Joint-Training. Both methods

TABLE VI
PERFORMANCE OF DIFFERENT STRATEGIES FOR AGGREGATING PROMPT BASES ON THE FIRST CISR BENCHMARK

Method Variants	PSNR	SSIM	FRMSE
nearest neighbor	28.00	0.793	0.0193
weighted combination	**28.57**	**0.805**	**0.0098**

TABLE VII
PERFORMANCE OF DIFFERENT STRATEGIES FOR USING PROMPTS TO TUNE THE BACKBONE MODEL ON THE FIRST CISR BENCHMARK

Method Variants	PSNR	SSIM
using prompts as prefix tokens	28.40	0.800
using prompts as adapters	**28.57**	**0.805**

use SwinIR as the backbone. We then evaluate BSRGAN [31] and SwinIR [6], pre-trained on DIV2K with complex and bicubic degradation. Table I reveals their limited performance, with large gaps from Individual-Training and Joint-Training due to dataset and degradation mismatch.

With SwinIR as the backbone model, we use sequential fine-tuning or LoRa to adapt its parameters, but these methods fail to retain prior task knowledge and offer limited improvement over the pre-trained SwinIR. The second and third benchmarks show that dataset and degradation changes individually affect SR performance, with degradation changes causing more severe forgetting than dataset changes. The first benchmark suggests their combination has a compounded effect, leading to a higher forgetting rate.

We evaluate existing continual learning algorithms, including L2P [2], DualP [41], CODA-P [3], and Dap-CL [1], by replacing their classification backbone models with SwinIR. These methods struggle with catastrophic forgetting in CISR, with most FRMSE values over 0.2 on the first and second benchmarks. L2P may disrupt learned prompts, compromising its ability to retain past knowledge. DualP and CODA-P struggle with capturing dataset feature distributions due to key definition issues, leading to low task localization accuracy in testing. Similarly, DAP-CL adopts a prompt-querying mechanism akin to DualP to provide conditions for the prompt generator. Consequently, these methods perform worse than Seq-FT and LoRa, hindered by their limited adaptation capacity and their inability to prevent catastrophic forgetting.

Our method achieves consistently superior results compared to the pre-trained models, Seq-FT, LoRa, and existing continual learning methods. For example, our method outperforms Seq-FT by 0.70 and 1.27 PSNR on the first benchmark for $2\times$ and $4\times$ super-resolution.

D. Ablation Study

We perform detailed experiments to validate the impact of principal modules and discuss the effects of varying key hyper-parameter, on the first CISR benchmark.

1) Effectiveness of Principal Modules: As shown in Table II, we assess the effectiveness of key components including: (1) using the CLIP image encoder to guide the construction of prompt adapters denoted as 'CLIP', (2) using the degradation encoder to guide the construction of prompt adapters denoted as 'Degradation', (3) using a single prompt base and identity denoted as 'Single', and (4) using multi-granularity prompt bases and identities denoted as 'Multi'.

No.1 is the baseline implemented by sequentially fine-tuning the pre-trained model on five tasks. Our investigation reveals that prompt adapters guided by the CLIP image encoder (No.3) significantly enhance performance, improving PSNR by 0.31 and reducing FRMSE to 0.0024 compared to the baseline. This improvement stems from the CLIP image features' capability to discern different tasks and the rich prompt information provided by prompt bases, effectively balancing task adaptation and knowledge retention. Similarly, the degradation encoder-guided prompt adapters (No.5) also demonstrate substantial metric improvements, confirming the utility of degradation features in characterizing cross-domain differences.

Notably, using both encoders together (No.6) results in better performance than using them individually, as they facilitate the learning of complementary prompt bases that address image domain and degradation type variances. Conversely, configurations employing only a single base and identity (No.2,4,7) exhibit performance degradation in both PSNR and SSIM metrics. This demonstrates that multi-granularity bases and identities help provide more effective prompt information due to the modeling of intra-task diversity.

2) Designs of Prompt Bases: To verify that shared down-projection $\mathbf{U}$ facilitates the capture of task-level prompt information, we compared the performance of using the separate up-projection ($\mathbf{U}$) component for constructing prompt bases. As shown in Table III, our shared $\mathbf{U}$ improves the PSNR by 0.57 compared to using a separate $\mathbf{U}$, indicating that the shared down-projection enhances the consistency and efficiency of task-level feature extraction.

3) Strategies for Combining Prompt Bases: We attempt to select the prompt bases having the nearest distances to the query features to construct the prompt adapter, labeled 'nearest neighbor' in Table VI. It underperforms the weighted combination by 0.53 PSNR and 0.0124 SSIM, due to the latter's robustness against mis-selected prompt bases and effectiveness in making full use of prompt bases across tasks. Furthermore, the weighted method offers better generalization, making it more suitable for broader use cases.

4) Distance Metrics for Similarity Estimation: We further investigate the choice of distance metrics in the similarity estimation module for CLIP features and degradation features. As shown in Table IV, adopting cosine distance for CLIP features and Euclidean distance for degradation features achieves the best overall performance. This result validates our metric-aligned design. The reason is that CLIP is pretrained with a contrastive objective based on cosine similarity, while Euclidean distance is more consistent with the MSE-trained degradation feature space.

5) Generality Across Different Backbone Architectures: To demonstrate the robustness of our LPA method, we conduct experiments on three different backbone architectures, namely DAT [21], HAT [54], and SwinIR [6], with the results presented in Table V. Compared to the sequential fine-tuning

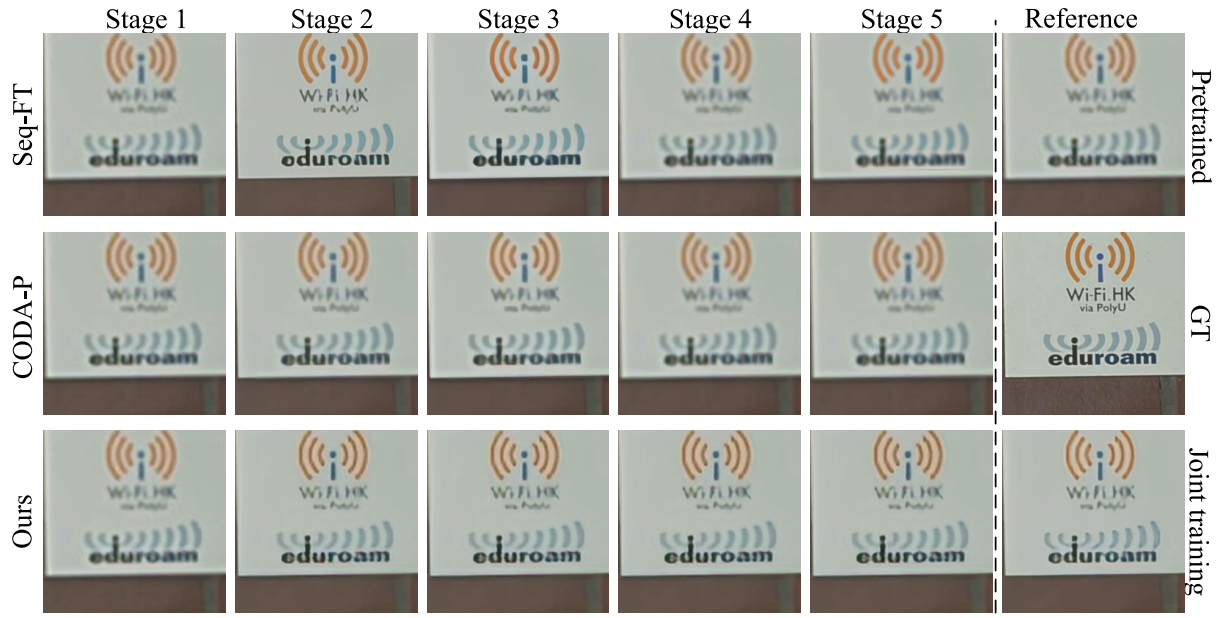

Fig. 5. Visualization of SR results generated by Seq-FT, CODA-P, and our method across different training stages. The example is selected out from the second task of the first CISR benchmark. The last column presents result of pre-trained SwinIR, ground-truth (GT), and result of Joint-Training, from top to bottom. Our method can preserve the ability to generate high-quality results in later stages, while the results of Seq-FT and CODA-P deteriorate severely.

(Seq-FT) baseline, our LPA method consistently improves continual learning performance across all three architectures, yielding average PSNR/SSIM gains of 0.74 dB/0.006 on DAT, 0.86 dB/ 0.0013 on HAT, and 1.27 dB/0.016 on SwinIR.

Notably, LPA achieves substantial improvements for all backbones, indicating effective mitigation of catastrophic forgetting. These results demonstrate that LPA is robust to different backbone architectures, validating the generalizability of our prompt-based adaptation mechanism.

6) Strategies for Using Prompts to Tuning Backbone: We try to use the prompt to create prefix tokens for keys and values in Transformers, instead of layer adapters. As shown in Table VII, this method of adding to prefix tokens decreases PSNR by 0.17 and SSIM by 0.005 compared to using prompts as adapters, since adapters are more flexible in adapting pixel-level features than prefix tokens.

7) Choice of Hyper-Parameters: In Fig. 7, we show the performance of our method with different numbers of identities (K) and rank values of prompt bases (r). We investigated the feasibility of employing the elbow method by exploring different values of K. Increasing K refines the modeling of the query feature space and enriches the prompt bases, improving PSNR. The optimal K value, which aligns with that determined by the elbow method, is equal to 3. Meanwhile, the rank r adjusts the trainable parameters and knowledge capacity within the prompt bases. Increasing r from 3 to 18 boosts PSNR, stabilizing around 12.

8) Performance on Unseen Datasets: To evaluate the generalization capability of our method, we test Seq-FT, Joint-Training, DualP, CODA-P and our method on unseen datasets, including Urban100, Flickr2K, and DRealSR in which LR images are created with bicubic, complex, and real-world degradation, respectively. As presented in Table VIII, our method achieves superior performance to Seq-FT and Joint-Training. The reason is that full-parameter training methods tend to overfit all data distributions, resulting in limited generalization ability to unseen datasets. By selecting prompts learned from tasks similar to the target image SR task to adapt the backbone without modifying the original parameters, our method demonstrates superior performance on unseen datasets.

TABLE VIII

PERFORMANCE ON UNSEEN DATASETS: URBAN100, FLICKR, AND DREALSR, WHERE BICUBIC, COMPLEX, AND REAL-WORLD DEGRADATION IS USED FOR SIMULATING LR IMAGES, RESPECTIVELY

Method	Urban100-b		Flickr-c		DRealSR-r		Average	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Seq-FT	23.94	0.704	24.29	0.610	30.64	0.862	26.29	0.725
Joint training	24.82	0.737	24.95	0.661	30.47	0.867	26.75	0.755
Lora [46]	17.78	0.362	20.71	0.441	24.02	0.617	20.84	0.470
L2P [2]	24.07	0.709	24.32	0.616	30.65	0.863	26.35	0.730
DualP [41]	24.20	0.714	24.33	0.616	30.67	0.863	26.40	0.730
CODA-P [3]	24.63	0.732	24.29	0.613	30.64	0.863	26.52	0.740
Ours	**25.63**	**0.776**	**24.85**	**0.649**	**30.66**	**0.864**	**27.05**	**0.763**

E. Visualizations

1) Visualization of Inference Results Across Different Methods: In Fig. 6, we show SR results from FT (Fine-Tuning), CODA-P, SwinIR, and our method across five training stages of the first CISR setting. Each row of images corresponds to different degradation types: bicubic degradation, real-world degradation, complex degradation, manga, and JPEG compression. The SwinIR and FT methods exhibit significant catastrophic forgetting, as they tend to overfit to the current task, leading to poor generalization across tasks. CODA-P struggles to generalize across diverse degradation patterns,

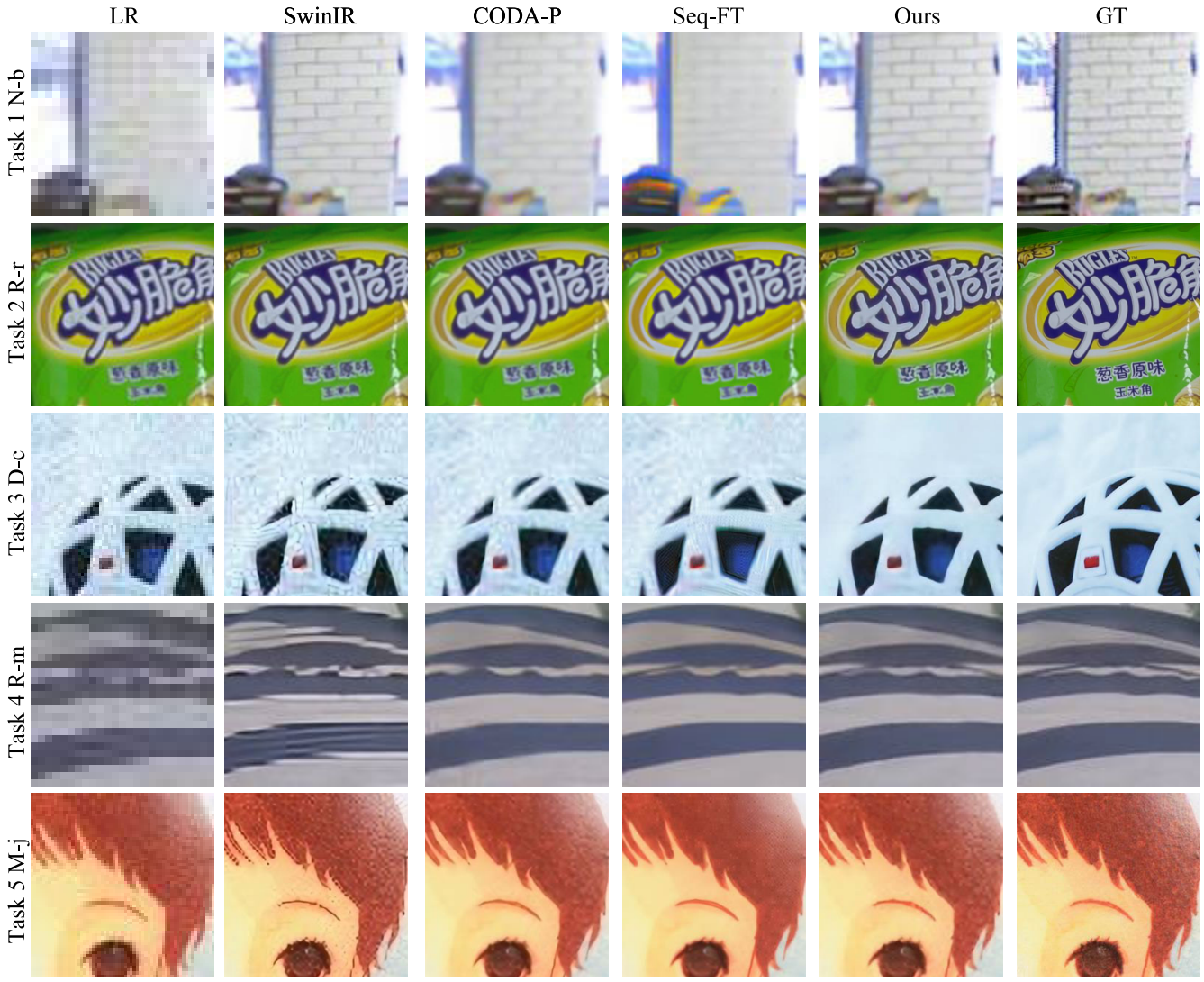

Fig. 6. Visualization of inference SR results on five tasks in the first experiment setting produced by pre-trained SwinIR (2nd column), Seq-FT (4th column), CODA-P (3rd column), and our method (5th column). The LR and GT images are in the 1st and 6th columns, respectively.

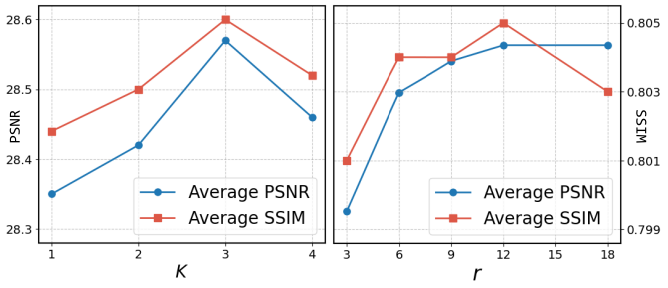

Fig. 7. Results of using different numbers of identities per task (K) and rank values for the prompt bases (r).

resulting in suboptimal texture clarity and detail restoration. In contrast, our method achieves higher restoration clarity and more precise texture recovery than the others.

2) Visualization of Results Comparing Different Methods Across Training Stages: In Fig. 5, we visualize the results of Seq-FT, CODA-P, and our method after each training stage, on one example from the second task of the first CISR benchmark. At the second training stage, all methods produce high-quality super-resolution results with clear appearance and sharp edges. However, as the training stage further increases, Seq-FT and CODA-P are unable to preserve the image SR ability of the second task and produce blurry results. Our method can preserve the knowledge of the second task stably, producing high-quality results after the training of subsequent tasks.

3) Visualization of Identity Features: We further visualize the features from the CLIP image and degradation encoders in Fig. 9. It can be observed that the CLIP image encoder has a strong ability to discriminate datasets, while the degradation encoder is rather advantageous at discriminating degradation types. These phenomena indicate that it is reasonable to use the two encoders to construct the prompt identities.

4) Visualization of PSNR Changes Over Training Stages: To better visualize forgetting behavior in our approach, we present the variation of PSNR across training stages. As shown in Fig. 8, the PSNR trends of different methods on each task of the first CISR benchmark demonstrate that the

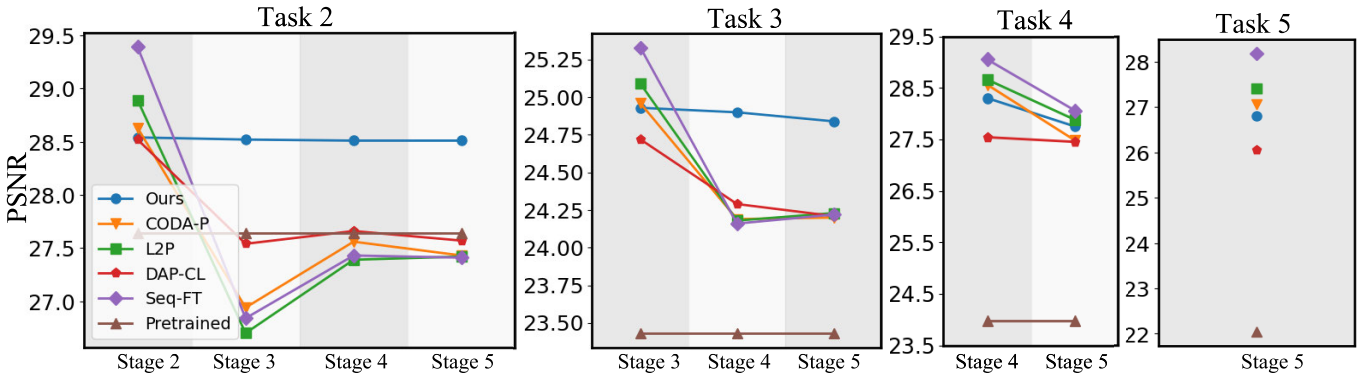

Fig. 8. Stage-wise forgetting curves. PSNR of each task across training stages on the first CISR benchmark. The x-axis denotes training stage and the y-axis denotes PSNR. Larger drops indicate more severe forgetting. Our method shows smoother trends than Seq-FT, DAP-CL, L2P, and CODA-P.

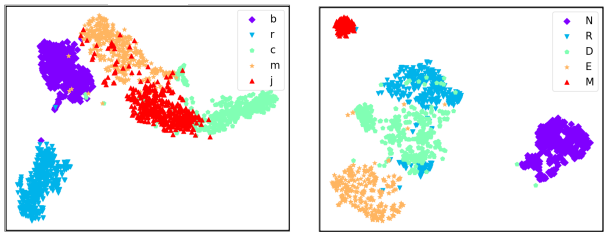

Fig. 9. The t-SNE visualization of features obtained from the degradation encoder (left) and CLIP image encoder (right).

pre-trained model cannot generalize well to different tasks. Seq-FT and existing continual learning methods (Dap-CL, L2P, CODA-P), struggle to balance the new task adaptation and history knowledge preservation abilities. Our method is more effective in alleviating the catastrophic forgetting issue while possessing adaptation ability comparable to L2P and CODA-P.

V. Conclusion

We pioneer a novel approach, *Learning Prompt Adapters*, to address the continual image super-resolution challenge. Our method introduces multi-granularity prompt bases to restore task-aware prompt information for each incoming SR task. We signify prompt bases with specific identities which help locate the target task and retrieve the relevant bases, ensuring the scalability to new tasks while alleviating the catastrophic forgetting issue. The utilization of prompt bases and identities enables the dynamic construction of prompt adapters that are used to tune the backbone model towards the target task. Experiments on three continual image SR settings, built upon NYU, RealSR, DIV2K, REDS, and MANGA109 datasets with various degradation types, demonstrate the superiority of our method over state-of-the-art methods.

References

[1] D. Jung, D. Han, J. Bang, and H. Song, "Generating instance-level prompts for rehearsal-free continual learning," in *Proc. ICCV*, 2023, pp. 11847–11857.

[2] Z. Wang et al., "Learning to prompt for continual learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 139–149.

[3] J. S. Smith et al., "CODA-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 11909–11919.

[4] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. ECCV*. Cham, Switzerland: Springer, 2014, pp. 184–199.

[5] 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. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3086–3095.

[6] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.

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

[8] M. Jia et al., "Visual prompt tuning," in *Proc. ECCV*. Cham, Switzerland: Springer, 2022, pp. 709–727.

[9] Z. Li, Y. Lei, C. Ma, J. Zhang, and H. Shan, "Prompt-in-prompt learning for universal image restoration," 2023, *arXiv:2312.05038*.

[10] B. Li et al., "PromptCIR: Blind compressed image restoration with prompt learning," 2024, *arXiv:2404.17433*.

[11] S. Khan, V. Potlapalli, F. Shahbaz Khan, and S. W. Zamir, "PromptIR: Prompting for all-in-one image restoration," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 71275–71293.

[12] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. ICML*, 2021, pp. 8748–8763.

[13] P. K. N. Silberman, D. Hoiem, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. ECCV*, 2012, pp. 746–760.

[14] E. Agustsson and R. Timofte, "NTIRE 2017 challenge on single image super-resolution: Dataset and study," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1122–1131.

[15] S. Nah et al., "NTIRE 2019 challenge on video deblurring and super-resolution: Dataset and study," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1996–2005.

[16] Y. Matsui et al., "Sketch-based Manga retrieval using Manga109 dataset," *Multimedia Tools Appl.*, vol. 76, no. 20, pp. 21811–21838, Oct. 2017.

[17] H. Chen, J. Gu, and Z. Zhang, "Attention in attention network for image super-resolution," 2021, *arXiv:2104.09497*.

[18] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[19] X. Zhang, H. Zeng, S. Guo, and L. Zhang, "Efficient long-range attention network for image super-resolution," in *Proc. ECCV*. Cham, Switzerland: Springer, 2022, pp. 649–667.

[20] Y. Zhou, Z. Li, C.-L. Guo, S. Bai, M.-M. Cheng, and Q. Hou, "SRFormer: Permuted self-attention for single image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 12734–12745.

[21] Z. Chen, Y. Zhang, J. Gu, L. Kong, X. Yang, and F. Yu, "Dual aggregation transformer for image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 12278–12287.

[22] X. Li, J. Dong, J. Tang, and J. Pan, "Dlgsanet: Lightweight dynamic local and global self-attention networks for image super-resolution," in *Proc. ICCV*, 2023, pp. 12792–12801.

[23] M. Protter, M. Elad, H. Takeda, and P. Milanfar, "Generalizing the nonlocal-means to super-resolution reconstruction," *IEEE Trans. Image Process.*, vol. 18, no. 1, pp. 36–51, Jan. 2009, doi: 10.1109/TIP.2008.2008067.

[24] T. Peleg and M. Elad, "A statistical prediction model based on sparse representations for single image super-resolution," *IEEE Trans. Image Process.*, vol. 23, no. 6, pp. 2569–2582, Jun. 2014, doi: 10.1109/TIP.2014.2305844.
[25] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010, doi: 10.1109/TIP.2010.2050625.
[26] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang, "Coupled dictionary training for image super-resolution," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3467–3478, Aug. 2012, doi: 10.1109/TIP.2012.2192127.
[27] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3262–3271.
[28] J. Gu, H. Lu, W. Zuo, and C. Dong, "Blind super-resolution with iterative kernel correction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1604–1613.
[29] S. Bell-Kligler, A. Shocher, and M. Irani, "Blind super-resolution kernel estimation using an internal-gan," in *Proc. NeurIPS*, 2019, pp. 284–293.
[30] Z. Yue, Q. Zhao, J. Xie, L. Zhang, D. Meng, and K.-Y.-K. Wong, "Blind image super-resolution with elaborate degradation modeling on noise and kernel," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2128–2138.
[31] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, "Designing a practical degradation model for deep blind image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4771–4780.
[32] X. Xu, D. Sun, J. Pan, Y. Zhang, H. Pfister, and M.-H. Yang, "Learning to super-resolve blurry face and text images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 251–260.
[33] S. Xi, J. Wei, and W. Zhang, "Pixel-guided dual-branch attention network for joint image deblurring and super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 532–540.
[34] J. Kirkpatrick et al., "Overcoming catastrophic forgetting in neural networks," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 13, pp. 3521–3526, 2017.
[35] F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence," in *Proc. ICML*, 2017, pp. 3987–3995.
[36] G. Zeng, Y. Chen, B. Cui, and S. Yu, "Continual learning of context-dependent processing in neural networks," *Nature Mach. Intell.*, vol. 1, no. 8, pp. 364–372, Aug. 2019.
[37] A. A. Rusu et al., "Progressive neural networks," 2016, *arXiv:1606.04671*.
[38] S. Yan, J. Xie, and X. He, "DER: Dynamically expandable representation for class incremental learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3013–3022.
[39] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "ICaRL: Incremental classifier and representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5533–5542.
[40] Y. Wu et al., "Large scale incremental learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 374–382.
[41] Z. Wang et al., "Dualprompt: Complementary prompting for rehearsal-free continual learning," in *Proc. ECCV*. Cham, Switzerland: Springer, 2022, pp. 631–648.
[42] Z. Liu, Y. Peng, and J. Zhou, "Insvp: Efficient instance visual prompting from image itself," in *Proc. ACM MM*, 2024, pp. 6443–6452.
[43] Z. Liu, X. Zou, G. Hua, and J. Zhou, "Token coordinated prompt attention is needed for visual prompting," 2025, *arXiv:2505.02406*.
[44] Y. Liu and M. Yang, "Sec-prompt: Semantic complementary prompting for few-shot class-incremental learning," in *Proc. CVPR*, Jun. 2025, pp. 25643–25656.
[45] J. He, Z. Duan, and F. Zhu, "CL-LoRA: Continual low-rank adaptation for rehearsal-free class-incremental learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 30534–30544.
[46] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," in *Proc. ICLR*, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9
[47] M. Le et al., "Mixture of experts meets prompt-based continual learning," in *Proc. NeurIPS*, vol. 37, 2024, pp. 119025–119062.
[48] W. Wang, H. Zhang, Z. Yuan, and C. Wang, "Unsupervised real-world super-resolution: A domain adaptation perspective," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4298–4307.
[49] X. Xu et al., "Dual adversarial adaptation for cross-device real-world image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5657–5666.
[50] Y. Liu, J. He, J. Gu, X. Kong, Y. Qiao, and C. Dong, "DegAE: A new pretraining paradigm for low-level vision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23292–23303.
[51] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proc. ACM-SIAM Symp. Discrete Algorithms*, 2007, pp. 1027–1035. [Online]. Available: https://api.semanticscholar.org/CorpusID:1782131
[52] A. Chaudhry et al., "Continual learning with tiny episodic memories," in *Proc. Workshop Multi-Task Lifelong Reinforcement Learn.*, 2019.
[53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[54] X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong, "Activating more pixels in image super-resolution transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22367–22377.

Chaowei Fang (Member, IEEE) received the Ph.D. degree from The University of Hong Kong in 2019. He is currently an Associate Professor with Xidian University. His research interests include multimodal modeling and robust learning.

Bolin Fu received the B.S. degree from Xidian University, Xi'an, China, in 2023, where he is currently pursuing the master's degree with the School of Artificial Intelligence. His research interests include multimodal modeling and image processing.

De Cheng received the B.S. and Ph.D. degrees from Xi'an Jiaotong University, China, in 2011 and 2017, respectively. He is currently an Associate Professor with the School of Telecommunications Engineering, Xidian University, China. His research interests include pattern recognition and machine learning.

Chengpei Tang received the Ph.D. degree from Sun Yat-sen University, China, in 2007. He is currently a Research Fellow with Sun Yat-sen University. His research interests include the Internet of Things and edge intelligence.

Guanbin Li (Member, IEEE) received the Ph.D. degree in computer science from The University of Hong Kong, Hong Kong, in 2016. He is currently an Associate Professor with the School of Computer Science and Engineering, Sun Yat-sen University, China. His research interests include computer vision and deep learning.