Task-aware All-in-one Guided Image Super-Resolution

Tingting Wang, Jun Wang, Qiuhai Yan, Junkang Zhang, Faming Fang, Guixu Zhang

***Abstract*—Guided image super-resolution (GISR) aims to enhance the resolution of a low-resolution (LR) target image by leveraging complementary information from a high-resolution (HR) guidance image. However, due to the substantial modality diversity among GISR subtasks, most existing methods are tailored to individual subtasks, which significantly limits their generalizability and practical scalability. To address this limitation, we propose MAG-Net, the first all-in-one GISR framework capable of handling multiple GISR subtasks within a single unified model. MAG-Net is built upon a shared encoder-decoder backbone and integrates two key modules to handle heterogeneous input modalities and diverse task objectives. The *Multi-modal Prompt Generation Module* (MPGM) dynamically generates task-aware prompts by jointly encoding the input image pair and pre-defined textual task descriptions. These prompts serve as soft instructions, effectively capturing both visual features and textural cues, thus enabling the model to adaptively distinguish and respond to different GISR subtasks. The *Multi-Guidance Routing Module* (MGRM) is then designed to mitigate task interference and enhance feature specialization. This module leverages a Mixture-of-Experts (MoE) strategy to adaptively route intermediate features through task-relevant expert branches, guided by the task-aware prompt and the characteristics of the guidance image. Extensive experiments across various GISR subtasks demonstrate that MAG-Net achieves state-of-the-art performance in both all-in-one and one-by-one training settings. Code and pre-trained models will be released upon paper acceptance.**

***Index Terms*—Guided image super-resolution, All-in-one, Pansharpening, MR image super-resolution, Depth image super-resolution**

I. Introduction

IMAGE super-resolution (SR) refers to the task of reconstructing a high-resolution (HR) image from its low-resolution (LR) counterpart, aiming to enhance visual quality and recover fine-grained structural details. Within the broader scope of SR, a specialized category known as guided image super-resolution (GISR) has garnered increasing attention. Unlike traditional single-image SR that relies solely on the degraded input, GISR introduces an auxiliary HR image from a different modality to guide the SR process. This guidance image, often referred to as prior information or an adjunctive cue, provides complementary structural or contextual information that facilitates more accurate and detailed reconstruction. By leveraging this additional guidance, GISR effectively mitigates the ill-posed nature of SR tasks and significantly improves the fidelity of the reconstructed results. The guidance image may manifest in various forms, such as constraints, priors, or reference images, and enables the model to intelligently infer high-frequency components that are otherwise challenging to recover from the LR input alone.

Typical subtasks under the GISR framework include pansharpening [1], [2], multi-contrast magnetic resonance (MR) image SR [3], [4], and depth image SR guided by corresponding RGB images [5]. Pansharpening, for instance, utilizes an HR panchromatic (PAN) image to guide the enhancement of an LR multispectral (LRMS) image, effectively overcoming the spatial-spectral resolution trade-off typically encountered in satellite imaging systems. In depth image SR, the HR RGB image provides fine texture and edge information to guide the enhancement of the LR depth image, enabling more precise recovery of geometric structures and improving visual coherence. For MR image SR, the resolution of a specific contrast in a target MR image is improved using guidance from an auxiliary contrast image, thereby enhancing the diagnostic value and anatomical clarity in medical imaging scenarios.

In recent years, deep learning based GISR have garnered significant attention. Methods founded on Convolutional Neural Networks (CNNs) [6], Transformers [7], and generative models such as diffusion [2] and Mamba [1] have been consistently introduced. However, due to significant cross-task gaps (e.g., variations in image modalities and application contexts) among various subtasks of GISR, the vast majority of methods only focus on one specific subtask, leading to the formulation of task-specific networks and loss functions, as illustrated in Figure 1(a). This specialization often results in limited model generalizability, impeding the transferability of models across diverse subtasks. Although some methods have attempted to design models that can be applied to multiple GISR subtasks [8], they still require training individual models for different subtasks, which limits the flexibility of the models. This predicament prompts a quest for a more adaptable GISR framework that can dynamically accommodate various SR scenarios with greater flexibility.

Recently, the concept of all-in-one frameworks has gained significant traction in the field of image restoration (IR), with a growing body of research exploring this direction [9], [10]. These frameworks are designed to employ a single unified model to address multiple IR tasks, such as image dehazing, denoising, and deraining. While these tasks typically involve images of the same modality, enabling the sharing of features and representations across tasks, the situation in GISR is considerably more complex. In GISR, the input usually consists of two images from different modalities. The heterogeneity between these modalities introduces a substantial domain gap, making it significantly more challenging to design a single network capable of generalizing across diverse GISR subtasks.

To empirically demonstrate this challenge, we employ a standard encoder-decoder backbone network to independently

Tingting Wang, Jun Wang, Qiuhai Yan, Junkang Zhang, Faming Fang and Guixu Zhang are with the School of Computer Science and Technology, East China Normal University, Shanghai, 200241, China. (e-mail: tingtingwang@cs.ecnu.edu.cn; 51265901080@stu.ecnu.edu.cn, yanqiuhai16@163.com, 52215901001@stu.ecnu.edu.cn; fmfang@cs.ecnu.edu.cn; gxzhang@cs.ecnu.edu.cn).

0000–0000/00$00.00 © 2021 IEEE

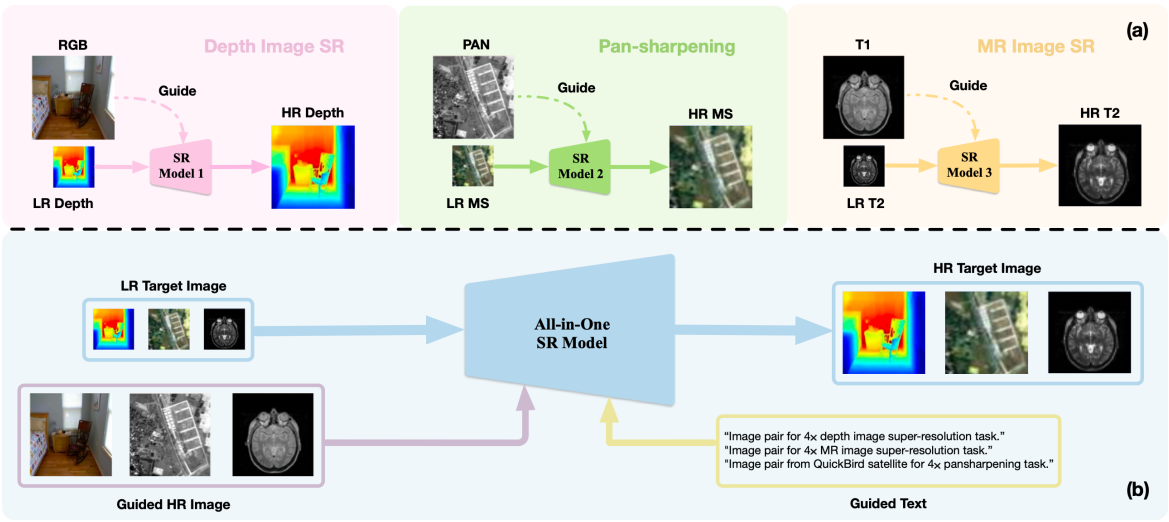

Fig. 1. (a) Task-specific SR networks designed for each GISR subtask; (b) Our proposed all-in-one SR network to handle various GISR subtasks simultaneously.

perform several GISR subtasks. The network accepts as input the concatenation of an upsampled LR target image and its corresponding HR guidance image, and outputs the SR target image. As illustrated in Figure 2, we analyze the L_2 norm distributions of feature maps across selected channels in the encoder stage. The results reveal a high degree of task-specific variation in feature activations, indicating that features extracted from different GISR subtasks exhibit significant independence. This observation suggests that when a unified network is used to handle multiple GISR subtasks, the variation in input modalities can cause interference among tasks, ultimately hampering the network's ability to achieve optimal performance across all subtasks. Based on this analysis, two critical challenges must be addressed in designing a unified GISR framework: *(a) how to effectively process and integrate input images from varying modalities; (b) how to mitigate or eliminate negative interference between different GISR subtasks within a single network architecture.*

To tackle the aforementioned challenges, we introduce a novel and unified framework for GISR, termed MAG-Net, as illustrated in Figure 1(b). MAG-Net is the first all-in-one GISR model that effectively integrates multiple GISR subtasks, including pansharpening, MR image SR, and depth image SR, within a single, flexible architecture. At the core of MAG-Net lies a shared encoder-decoder backbone, upon which we design two key modules: the Multi-modal Prompt Generation Module (MPGM) and the Multi-Guidance Routing Module (MGRM). The MPGM dynamically generates task-aware multi-modal prompts by jointly encoding the input image pair and pre-defined textual guidance that encapsulates task-specific description. These prompts act as implicit instructions, encoding both visual characteristics and textual information from the input images and task descriptions. By bridging image and text modalities, the prompts help disambiguate task-specific objectives and suppress cross-task interference, thereby enabling effective task differentiation within the unified model. To further leverage the guidance information encoded by the MPGM, we propose the MGRM,

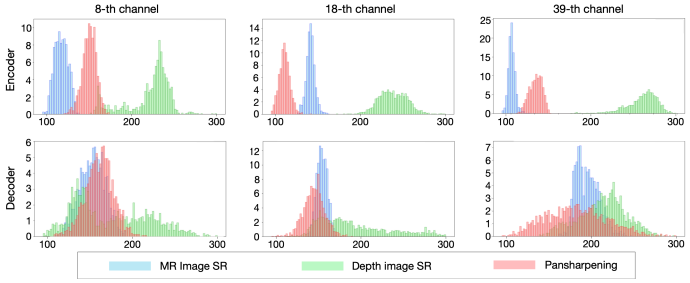

Fig. 2. The distribution of L_2 norm for the feature maps of some channels when using an encoder-decoder network architecture to handle different GISR subtasks. Note that each subtask is trained individually.

which adopts a Mixture-of-Experts (MoE) mechanism. This module dynamically routes intermediate features through specialized expert branches, selected adaptively based on the characteristics of the guided input. This enables MAG-Net to selectively activate and exploit the most relevant knowledge for each task, ensuring both generality and specialization.

The main contributions of our proposed method are concluded as follows.

1) **A unified all-in-one GISR framework:** We propose the first unified architecture that accommodates a diverse set of GISR subtasks. By bridging the modality and task gaps, MAG-Net offers a generalizable solution for cross-domain guided SR.
2) **A novel task-aware guidance mechanism:** Our design incorporates multi-modal prompt generation and adaptive feature routing through MoE, enabling precise, task-specific feature enhancement.
3) **Extensive evaluation with state-of-the-art results:** We conduct comprehensive experiments on representative GISR benchmarks, demonstrating that MAG-Net consistently achieves superior performance compared to existing methods.

II. Related Work

A. Guided image super-resolution

Compared with SISR, GISR incorporates an additional HR guidance image to improve the resolution of an LR target image by transferring structural information from the HR guidance image. A commonly adopted method in GISR is based on joint or guided filtering, where the filter leverages the guidance image as a prior to transmit structural cues. Several joint filtering approaches have been introduced [11], [12]. Nonetheless, when structural discrepancies exist between the HR guidance image and the target image, misalignment issues may occur. To handle this, an alternative method [13] constructs an explicit mapping function to model structural differences across different image modalities. Song et al. [14] proposed a GISR framework based on coupled dictionary learning, which learns modality-specific dictionaries in a sparse feature domain to reduce inconsistency between the target and guidance images. However, such methods typically rely on manually designed objective functions, which might not sufficiently represent the natural image priors.

With the advent of deep learning in computer vision, CNNs have been increasingly applied to GISR. For instance, Yang et al. [15] proposed the PanNet architecture, incorporating upsampled LRMS images into the output, and training the model in the high-pass filtering domain. Wang et al. [16] introduced a multiscale U-shaped CNN (MUCNN) that uses a specially designed spectral-to-spatial convolution (SSconv) for upsampling. These CNN-based methods are entirely data-driven and require a large volume of paired training examples.

To enable end-to-end learning for GISR, deep unrolling networks have been widely explored. Marivani et al. [17] introduced a multimodal unfolding network inspired by convolutional sparse coding. Deng et al. [18] proposed a joint multimodal dictionary learning (JMDL) method and later unfolded it via ISTA. Yang et al. [19] incorporated memory mechanisms and dual priors into a conditional unfolding framework. Lei et al. [20] designed MC-CDic under an optimization-guided fidelity framework. Zhou et al. [8] further combined MAP estimation with LSTM-based persistent memory to model structural similarities and reduce information loss.

B. All-in-one image restoration

Recently, all-in-one natural IR has emerged as a prominent approach, which aims to tackle various restoration tasks simultaneously by leveraging a single universal model.

Recent advances in all-in-one image restoration (IR) have leveraged contrastive learning, prompt-based paradigms, and multimodal priors to enhance generalization across diverse degradations. AirNet [9] pioneers blind all-in-one IR via contrastive degradation representation learning. Extending this, IDR [21] models degradations through physical priors in a two-stage framework. PromptIR [22] and PIP [23] introduce dynamic and nested prompt modules to boost restoration adaptivity, while Chen et al. [24] generalize prompt learning to 30+ tasks via a visual task prompt framework. MPerceiver [25] further explores multimodal prompting using generative priors from stable diffusion. GRIDS [26] proposes group-wise training and adaptive routing for multi-degradation IR. In medical imaging, AMIR [27] and ProCT [28] employ prompt routing and view-aware strategies for various modality-specific tasks.

Recent advances have also explored the integration of large vision-language models, with DA-CLIP [29] demonstrating how pretrained CLIP models can be adapted for universal IR, and Perceive-IR [30] utilizing DINO-v2-based guidance modules to enhance restoration quality through semantic prior. TextPromptIR [31] leveraged textual prompts to guide the restoration process through natural language instructions.

III. Proposed Method

To systematically and comprehensively introduce our proposed method, we first present the overall pipeline, followed by a detailed explanation of the two efficient modules: MPGM and MGRM.

A. Overall pipeline

As depicted in Figure 3 (a), the proposed MAG-Net takes as input an LR target image $I_l \in \mathbb{R}^{H\times W\times C_1}$ and an HR guidance image $G_h \in \mathbb{R}^{sH\times sW\times C_2}$, where H and W represent the height and width of the target image, s denotes the SR scaling factor, and C_1, C_2 indicate the number of channels in the target and guidance images, respectively. These inputs are jointly utilized to reconstruct a high-quality SR target image by leveraging both intra-modal and cross-modal information.

Inspired by Restormer [32], we adopt a 4-level hierarchical encoder-decoder architecture as the backbone of MAG-Net. Each level of the hierarchy contains a stack of Transformer Blocks, with the number of blocks gradually increasing from the top (shallow) to the bottom (deep) levels. This design ensures a favorable trade-off between representational power and computational efficiency, allowing deeper layers to capture more complex and fine-grained features. To address the issue of task interference in multi-task GISR settings, we embed a dedicated **Multi-Guidance Routing Module (MGRM)** at each encoder level. The MGRM adaptively routes the intermediate features through specialized processing pathways according to task-specific guidance, thereby enabling more accurate and discriminative feature learning across heterogeneous GISR tasks. To further enhance the guidance capability of MGRM, we propose the **Multi-modal Prompt Generation Module (MPGM)**. This module dynamically generates a task-aware Multi-Modal Prompt (MMP) by jointly analyzing the paired input images and integrating pre-defined task-relevant textual descriptions. The MMP encodes both task-specific visual cues and predefined textural descriptions, acting as explicit instructions that inform the routing decisions made by the MGRM. By injecting the MMP into each level of the encoder, the model is better equipped to distinguish among different GISR subtasks and allocate features to specialized paths accordingly, thereby improving the overall adaptability and performance of the network in diverse GISR scenarios.

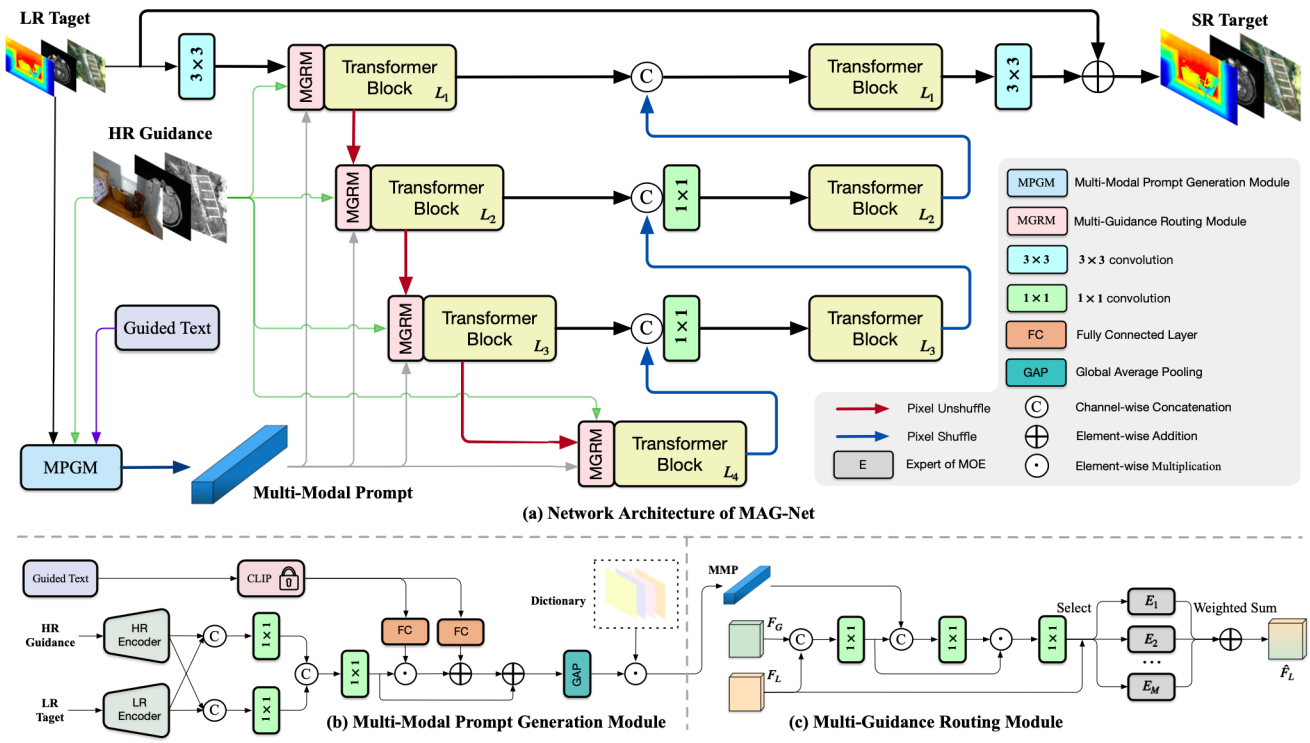

Fig. 3. (a) Overview of the proposed MAG-Net; (b) detailed architecture of the Multi-modal Prompt Generation Module (MPGM); (c) detailed architecture of the Multi-Guidance Routing Module (MGRM).

B. Multi-modal prompt generation module

As illustrated in Figure 3(a), we first extract shallow features from the two input images using separate encoders consisting of three convolutional layers, denoted as $E_H(\cdot)$ and $E_L(\cdot)$, respectively. These features are then fused and refined through task-specific branches to produce modality-aware intermediate representations. The generation of HR and LR features, denoted as F_H and F_L, can be formulated as:

$$F_H = \mathrm{Conv}_{1\times1}([\mathrm{E_H}(G_h); \mathrm{E_L}(I_l)]), \tag{1}$$

$$F_L = \mathrm{Conv}_{1\times1}([\mathrm{E_H}(G_h); \mathrm{E_L}(I_l)]), \tag{2}$$

where $[\cdot;\cdot]$ denotes the channel-wise concatenation operation. The concatenated features are then refined by a 1×1 convolution layer followed by a non-linear activation function (e.g., ReLU). These blocks are designed to filter and project the concatenated features into task-relevant embedding spaces. To further integrate information from both the HR and LR branches, we construct a fused representation F_I by aggregating F_H and F_L through an additional 1×1 convolution:

$$F_I = \mathrm{Conv}_{1\times1}([F_H; F_L]), \tag{3}$$

This fusion ensures that both high-frequency details from the HR image and structural cues from the LR image are effectively combined to serve as a rich visual prompt foundation.

Unlike natural image restoration tasks, GISR tasks exhibit greater heterogeneity, not only in terms of image modality (e.g., remote sensing, medicine, depth) but also in spatial resolution ratios and input channel dimensions. This diversity introduces challenges for a unified model to correctly differentiate and specialize its behavior for each task. Therefore, it becomes essential to inject more comprehensive, task-specific guidance into the network. Inspired by the role of textual instructions in human-computer interaction, and the success of multi-modal models such as CLIP [33], we incorporate textual task descriptions to augment the visual features. For each GISR subtask, we manually design concise, guided textual prompts that encode relevant information such as task type, SR ratio, or sensor characteristics (e.g., satellite type in pansharpening). Examples of these task descriptions are illustrated in Figure 1. By feeding these textual descriptions into a pre-trained CLIP text encoder, we obtain a task-aware text embedding F_T that complements the fused visual representation F_I. We then integrate the visual and textual information by feature modulation and skip connection as follows:

$$F_{IT} = F_I + (F_I \odot \mathrm{FC}_1(F_T) + \mathrm{FC}_2(F_T)), \tag{4}$$

where $\odot$ denotes the Hadamard product, enabling fine-grained interaction between the visual and textual modalities, FC_1 and FC_2 are two fully connected layers, respectively.

Finally, we use the joint representation F_{IT} to generate a set of weights for dynamically combining elements from a learnable dictionary. The resulting MMP, denoted as P_{mm}, is computed as:

$$P_{mm} = \sum_{i=1}^{N} w_i D_i, \qquad w_i = \mathrm{Softmax}(\mathrm{GAP}(F_{IT})), \tag{5}$$

where $\mathrm{GAP}(\cdot)$ represents global average pooling, and $D = [D_1, D_2, \ldots, D_N]$ is a shared, learnable dictionary designed to encode a set of reusable, task-agnostic feature atoms. The

generated MMP P_{mm} acts as a task-specific prompt that effectively guides downstream modules such as MGRM.

C. Multi-guidance routing module

To mitigate task interference among multiple GISR subtasks, we aim to allocate conflicting representations to distinct parameter subspaces, without introducing significant computational or memory overhead. *Mixture-of-Experts* (MoE) [34] provides an elegant mechanism for this by dynamically selecting expert subnetworks for each input, thereby enhancing model capacity and specialization while keeping the overall cost manageable. However, conventional MoE architectures typically rely solely on input token representations for routing decisions, neglecting crucial task-specific contextual signals. In the context of GISR, such an approach is suboptimal because the inputs span diverse modalities, each with unique structural and semantic characteristics. Furthermore, the HR guidance image G_h contains auxiliary information that is essential for interpreting and differentiating tasks.

To address these limitations, we enhance the routing mechanism by incorporating both the learned task-specific MMP P_{mm} produced by MPGM and the guided features F_G extracted from G_h,

$$F_G = \mathrm{E_G}(G_h), \tag{6}$$

where $E_G(\cdot)$ denotes an shallow encoder consisting of three convolutional layer. This enriched guidance enables the routing function to make more informed and discriminative decisions. The routing procedure, as illustrated in Figure 3(b), is formalized as follows:

$$F' = \mathrm{Conv}_{1\times1}([F_L; F_G]), \tag{7}$$

$$F_{\text{gate}} = \mathrm{Conv}_{1\times1}(F' \odot \mathrm{Conv}_{1\times1}([F'; P_{mm}])), \tag{8}$$

where F_L denotes the intermediate feature from L-th encoder level, and $\odot$ represents the Hadamard product. The 1×1 convolution layers serve as channel-wise fusion operations, integrating task-relevant and spatially-aware information into the gating signal F_{gate}.

Next, we use F_{gate} to compute the expert selection weights and perform a sparse, weighted aggregation of the outputs from a pool of experts to generate the enhanced feature $\hat{F}_L$:

$$\hat{F}_L = \sum_{j=1}^{M} \mathrm{K}(F_{\text{gate}})_j \cdot \mathrm{E}_j(F_L), \tag{9}$$

where $\mathrm{K}(\cdot)$ denotes a Top-K gating function that selects the K most relevant experts based on F_{gate}, setting the remaining weights to zero to ensure sparsity. In our implementation, M and K are set to 4 and 2, respectively. Each expert E_j is implemented as a lightweight multilayer perceptron (MLP) that transforms F_L into a task-specialized representation.

To improve load balancing and avoid over-reliance on a small subset of experts, we incorporate auxiliary load balancing loss terms $L_{Balance}$. In addition, to further reduce computational burden and minimize potential interference among experts, we adopt a distributed processing strategy, where each selected expert operates on a portion of F_L rather than the entire feature map.

D. Loss function

The total loss of the proposed MAG-Net can be expressed as:

$$L = L_1 + \gamma L_{Balance}, \tag{10}$$

where L_1 represents the reconstruction loss, measuring the difference between the reconstructed HR image and the ground truth (GT); $L_{Balance}$ is the regularization component designed for MoE to ensuring that the same small group of experts is not repeatedly selected. The parameter γ serves as a balancing weight.

IV. Experiments

To validate the effectiveness of the proposed MAG-Net, we conduct comprehensive experiments on three common GISR subtasks, i.e, pansharpening, depth image SR and MR image SR, from both quantitative and qualitative perspectives.

A. Datasets and evaluation metrics

For pansharpening task, we conduct experiments on datasets from three satellites: QuickBird (QB), WorldView-4 (WV4), and GaoFen-1 (GF1) [35]. To facilitate performance evaluation, we crop the original image pairs into small samples. Each training pair contains a 128×128×1 PAN image, a 32×32×4 LRMS image as well as a 128×128×4 GT (HRMS) image. Corresponding image quality assessment (IQA) metrics include the relative dimensionless global error in synthesis (ERGAS), peak signal-to-noise ratio (PSNR), and spectral angle mapper(SAM). For depth image SR task, we select widely-recognized benchmark NYU v2 [36] dataset, which contains 1449 pairs of RGB-D images. Following the settings of [37], we use 1000 pairs for training and the remaining 449 pairs for testing, with SR ratio setting to 4, 8, and 16, respectively. The RMSE metric is used to quantify the difference between the SR image and the GT image. For MR image SR task, we conduct evaluations with three different SR ratio on the BraTS dataset, which includes 285 multi-contrast MR volumes. For each volume, we select the middle 100 slices and exclude slices with insufficient content information. In total, we use 20,480 slices for training and 2,320 slices for testing. Besides, we evaluate the results using peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) to measure the quality of the SR results.

B. Training details

All experiments are conducted using PyTorch on a single NVIDIA RTX 3090 GPU. We use the Adam optimizer (batch size 4) with an initial learning rate of 4e-4, decayed via cosine annealing. The encoder-decoder backbone includes Transformer blocks with depths $L_1 = 3$, $L_2 = 4$, $L_3 = 4$, and $L_4 = 5$. In the all-in-one setting, the model is jointly trained on multiple datasets, with epoch length matched to the smallest dataset (1000 images for depth SR). Datasets are reshuffled after each epoch, and training runs for 500 epochs, including ablations. In the one-by-one setting, the model is trained separately for each task, each for 500 epochs.

C. Evaluations under all-in-one setting

Since we are the first to propose the all-in-one approach for the GISR task, there are no directly comparable methods. Therefore, we adapt several popular all-in-one IR methods to fit the GISR task. The selected comparison methods include: PromptIR [22], Gridformer [10], Transweather [38], CAPT-Net [39] and AdaIR [40]. Specifically, we concatenate the upsampled LR image with the guidance image as the input for these methods, and the output is the corresponding SR result. Notably, MAG-Net achieves the smallest model size among all benchmark methods, highlighting its efficiency in parameter usage.

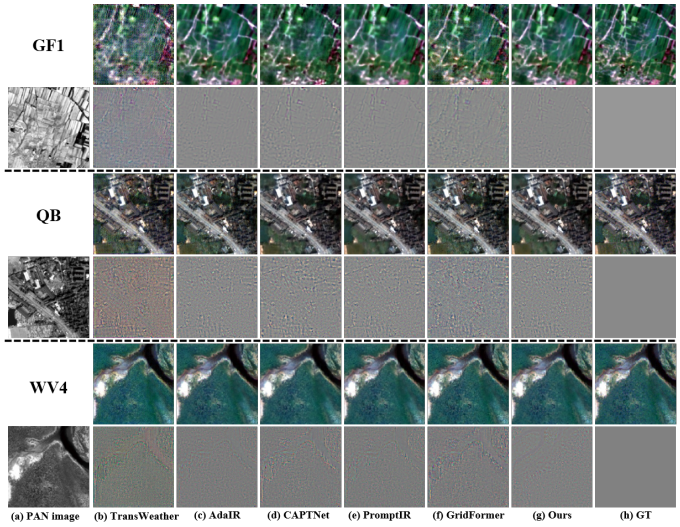

Fig. 4. Visual comparison and error maps of pansharpening under all-in-one setting.

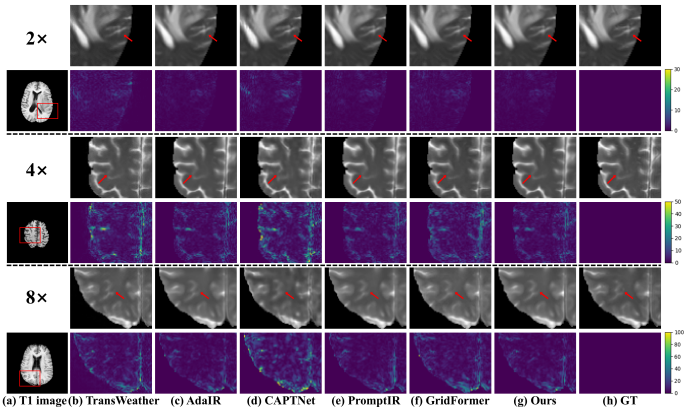

Fig. 5. Visual comparison and error maps of MR image SR under all-in-one setting.

The quantitative results of all-in-one approaches are reported in Table I. As shown in the table, our method consistently outperforms all competing methods across various tasks and settings, demonstrating superior performance on all evaluation metrics. For instance, on the WV4 dataset for the pansharpening task, our method achieves a PSNR that is 0.55 dB higher than the second-best method, AdaIR, indicating a substantial improvement in reconstruction fidelity. Among the

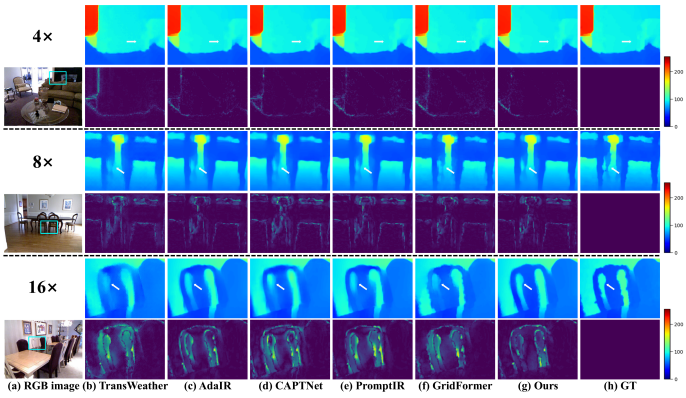

Fig. 6. Visual results and error maps of depth image SR under all-in-one setting.

baseline methods, AdaIR and PromptIR generally occupy the second and third ranks, respectively. However, both methods exhibit marked performance imbalances across different tasks and datasets. For example, although AdaIR performs competitively on the WV4 and QB datasets in the pansharpening task, it falls behind PromptIR on the GF1 dataset, suggesting that its generalization ability across domains is limited. This inconsistency underscores the difficulty of designing unified models that are equally effective across diverse task distributions. In contrast, the remaining three methods show significantly inferior performance compared to ours in nearly all subtasks and datasets. These results further validate the effectiveness and robustness of our proposed approach, particularly in multi-task scenarios where maintaining consistent performance across heterogeneous data sources is critical.

Qualitative comparisons are illustrated in Figures 4 to 6, where we not only display the SR results of all competing methods but also provide corresponding error maps, either for full images or zoomed-in regions of interest. Overall, our proposed MAG-Net consistently produces SR results that are visually closer to the GT images, with significantly better preservation of fine details and fewer artifacts. Specifically, in the pansharpening task shown in Figure 4, the error maps reveal that TransWeather suffers from severe spectral distortion, particularly in QB and WV4 cases. Other methods also exhibit varying degrees of spatial or spectral inconsistencies, indicating limitations in their ability to preserve both structure and spectrum. In the MR image SR task presented in Figure 5, CAPTNet shows noticeable loss of anatomical details, such as blurred tissue boundaries and weakened contrast. Note that AdaIR performs relatively better in this task, possibly due to its use of frequency-domain operations that are more compatible with the Fourier nature of k-space MRI data. For the depth image SR task in Figure 6, our MAG-Net clearly excels, especially around challenging object boundaries such as thin structures or occlusion edges. MAG-Net demonstrates a strong capability to accurately recover depth discontinuities and fine structures. This improvement can be largely attributed to our effective integration of the guidance image, which allows the network to better align structural cues and preserve task-relevant spatial information during SR process.

TABLE I
QUANTITATIVE RESULTS UNDER ALL-IN-ONE SETTING ACROSS THREE TASKS. BEST RESULTS ARE IN **BOLD**, AND SECOND-BEST RESULTS ARE UNDERLINED.

Method	Pansharpening									MR Image SR						Depth Image SR		
	WV4			QB			GF1			2×		4×		8×		4×	8×	16×
	PSNR↑	SAM↓	ERGAS↓	PSNR↑	SAM↓	ERGAS↓	PSNR↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	RMSE↓	RMSE↓	RMSE↓
Gridformer	40.57	1.42	1.19	47.72	0.96	0.70	45.59	0.98	0.97	42.15	0.9890	36.70	0.9662	33.64	0.9430	1.67	3.04	5.72
Transweather	39.54	1.67	1.39	46.26	1.14	0.80	43.99	1.22	1.18	39.91	0.9821	36.25	0.9594	33.73	0.9332	2.97	4.14	6.20
CAPTNet	41.10	1.38	1.15	47.92	0.94	0.68	48.94	0.70	0.77	40.43	0.9872	35.05	0.9573	32.09	0.9249	1.64	2.88	5.13
AdaIR	43.46	1.06	0.85	49.90	0.75	0.54	51.80	0.47	0.54	44.27	**0.9928**	38.92	0.9780	35.56	0.9585	1.53	2.79	5.09
PromptIR	43.34	1.08	0.88	49.76	0.76	0.55	52.20	0.46	0.53	44.16	0.9926	38.97	0.9781	35.58	0.9582	1.48	2.69	4.89
MAG-Net	**44.01**	**0.99**	**0.81**	**50.15**	**0.74**	**0.53**	**52.52**	**0.44**	**0.52**	**44.64**	**0.9928**	**39.21**	**0.9786**	**35.60**	**0.9679**	**1.29**	**2.47**	**4.45**

D. Evaluations under one-by-one setting

We further compare the proposed MAG-Net with SOTA single-task methods across three representative tasks. The quantitative results are summarized in table II, and the corresponding qualitative comparisons are illustrated in Figures 7 to 9.

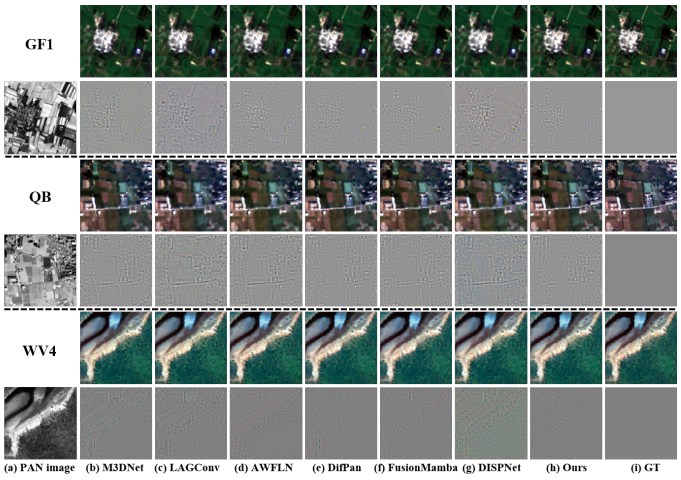

Fig. 7. Visual comparison results and error maps of pansharpening under one-by-one setting.

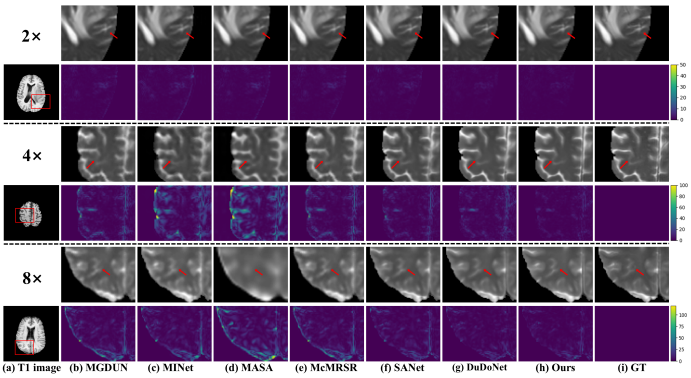

Fig. 8. Visual comparison results and error maps of MR image SR under one-by-one setting.

a) Pansharpening: For this task, we include the following strong baselines: AWFLN [6], DISPNet [41], LAGConv [42], M3DNet [43], FusionMamba [1], and Dif-Pan [2].

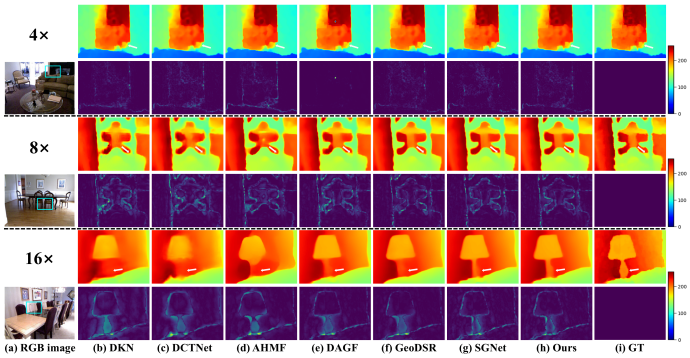

Fig. 9. Visual comparison results and error maps of depth image SR under one-by-one setting.

As shown in Table II, our MAG-Net and Dif-Pan consistently outperform other methods across the three benchmark datasets. While MAG-Net slightly lags behind Dif-Pan on the QB and GF1 datasets, it delivers competitive results on WV4 and even surpasses other approaches by a large margin in several metrics. Figure 7 presents the visual comparisons. The error maps clearly demonstrate that DISPNet and LAGConv suffer from severe spectral distortions and fail to recover spatial details effectively. Although M3DNet and AWFLN maintain relatively good spectral consistency, they still lose considerable structural information. In contrast, our MAG-Net achieves a better balance between spatial and spectral fidelity, resulting in visually superior pansharpened outputs with fewer artifacts and sharper textures.

b) MR image SR: The single-task baselines considered for this task include MGDUN [3], DuDoNet [4], MINet [44], MASA [45], McMRSR [7], and SANet [46]. According to Table II, our MAG-Net achieves the best performance across all three SR scales, demonstrating its strong ability to restore high-quality anatomical details from LR MR images. Visual comparisons in Figure 8 further verify these findings. Among all methods, our MAG-Net and DuDoNet yield the most accurate results compared to the ground truth. In particular, MAG-Net produces sharper tissue boundaries and better preserves subtle structures, offering slight but consistent improvements over DuDoNet.

c) Depth image SR: We compare MAG-Net with several SOTA depth SR methods, including GeoDSR [47], DAGF [48], AHMF [49], DCTNet [50], DKN [37], and

SGNet [5]. As reported in Table II, our MAG-Net ranks second only to SGNet in the 4× SR setting, while achieving the best performance for more challenging 8× and 16× scenarios. Figure 9 provides a visual comparison for the depth image SR task. It is evident that MAG-Net produces the most faithful restorations, especially around complex object boundaries. The reconstructed depth maps are more accurate and exhibit fewer artifacts, highlighting the advantages of our model in capturing fine geometric details.

E. Ablation studies and discussions

Our ablation experiments are conducted simultaneously on multiple GISR tasks. Due to space limitations, one IQA metric is selected for each task to measure the SR performance. The pansharpening and MR image SR tasks both use the PSNR metric, while the depth image SR task use the RMSE metric.

a) Impact of the two key modules: To further validate the effectiveness of the two core components, i.e., MPGM and MGRM, we conduct an ablation study under identical experimental settings. Specifically, we compare four configurations: (1) a baseline where both modules are removed, (2) a variant with MGRM removed, (3) a variant with MPGM removed, and (4) the full MAG-Net model with both modules enabled. The corresponding results are summarized in Table III.

From the comparison, it is evident that each component plays a vital role in improving the performance of the overall model. Removing either MPGM or MGRM leads to a noticeable performance drop, while removing both results in the most significant degradation. This demonstrates that the combination of MPGM and MGRM allows MAG-Net to dynamically select task-specific feature processing paths based on the characteristics of the source images and the task description. By enabling this adaptive routing mechanism, the model effectively reduces task interference and promotes better multi-task generalization.

To further support this conclusion, we visualize the task prompts generated by the MPGM using t-SNE, as shown in Figure 10. Figure 10(a) illustrates the prompts from all GISR subtasks across all datasets. The visualization reveals that the learned prompts form well-separated clusters corresponding to different tasks, indicating that the MPGM effectively captures discriminative and task-aware representations. Although the inter-task variation appears less pronounced in Figure 10(a), Figure 10(b) shows that prompts from the MR image SR task under three different SR ratios still exhibit clear distinctions. The reason lies in that these intra-task differences are smaller compared to the separability between different tasks. These prompt embeddings subsequently serve as meaningful guidance for the MGRM, enabling it to route features through the most suitable expert branches for each specific task and dataset.

b) Impact of task combinations: We conduct experiments by training our proposed MAG-Net on various combinations of GISR tasks to assess its robustness in multi-task learning scenarios. The quantitative results are reported in Table IV. Notably, even as the number of simultaneously learned tasks increases, MAG-Net maintains competitive performance without suffering from significant degradation. This robustness can

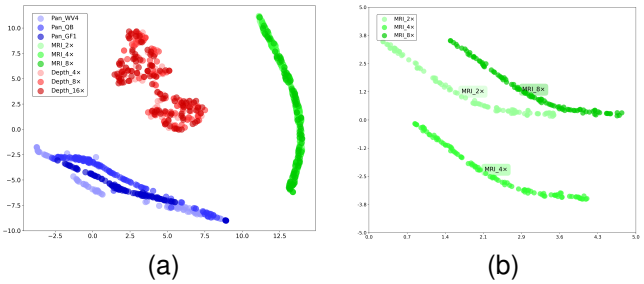

Fig. 10. t-SNE visualization of MMP from (a) different tasks; (b) different SR ratios for MR image SR task

be largely attributed to our multi-modal prompting strategy and the task-adaptive routing mechanism, both of which are explicitly designed to reduce task interference by encouraging the separation of task-relevant features.

Interestingly, in several task combination settings, the model even surpasses the performance of the corresponding single-task versions. This suggests that MAG-Net is not only effective in isolating cross-task interference, but also capable of leveraging complementary information across tasks to achieve mutual enhancement.

c) Impact of dataset diversity: In our experiments, for each GISR subtask, we select three different datasets or SR ratios to train the model. For the sake of clarity, we treat different SR ratios as distinct datasets. To investigate the impact of the number of datasets per task on model performance, we train MAG-Net using one or two datasets for each subtask, and report the results in Table V. From the results, we observe a consistent trend: as the number of datasets increases, the SR performance across nearly all tasks improves accordingly. This demonstrates that including more diverse datasets within the same subtask category can effectively enrich the model's representation capability. This finding also suggests that the domain gap among different datasets for the same task is relatively small, especially when compared to the domain gap across different tasks. This observation is further supported by the t-SNE visualization in Figure 10, where samples from the same task but different datasets tend to form overlapping clusters, indicating shared structural and spectral characteristics. Therefore, incorporating additional datasets does not introduce significant learning interference. On the contrary, it enables the model to better learn robust, task-relevant features, ultimately leading to superior SR performance across diverse real-world scenarios.

d) Impact of guiding modality: In our framework, the generation of multi-modal prompts leverages not only the visual information extracted from the guidance images but also the textual information embedded in the task-specific descriptions. To assess the individual and joint contributions of these two guiding modalities, we conduct an ablation study, and the results are summarized in Table VI. It is evident that using only a single modality or no guiding information significantly limits the model's ability to differentiate among various GISR tasks. In such cases, the generated prompts tend to exhibit a bias toward certain tasks, resulting in performance

TABLE II
ONE-BY-ONE GISR METRIC RESULTS ACROSS THREE TASKS. BEST PERFORMANCES ARE **BOLDED** AND SECOND BEST ARE UNDERLINED.

Pansharpening										MR Image SR							Depth Image SR			
Method	WV4			QB			GF1			Method	2×		4×		8×		Method	4×	8×	16×
	PSNR	SAM	ERGAS	PSNR	SAM	ERGAS	PSNR	SAM	ERGAS		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM		RMSE	RMSE	RMSE
AWFLN	42.13	1.23	0.98	49.14	0.82	0.60	49.70	0.62	0.64	MGDUN	40.97	0.9893	35.34	0.9632	32.61	0.9361	GeoDSR	1.42	2.62	4.86
DISPNet	40.79	1.43	1.15	48.18	0.93	0.70	46.61	0.86	0.84	MINet	38.84	0.9788	35.62	0.9625	32.38	0.9335	DAGF	1.36	2.87	6.06
LAGConv	41.54	1.33	1.07	47.89	0.96	0.71	47.66	0.79	0.79	MASA	40.37	0.9883	34.93	0.9590	30.41	0.8954	AHMF	1.40	2.89	5.64
M3DNet	42.27	1.20	0.97	49.45	0.80	0.58	49.80	0.62	0.64	McMRSR	40.96	0.9893	35.20	0.9596	30.55	0.8985	DCTNet	1.59	3.16	5.84
FusionMamba	42.84	1.15	0.92	49.50	0.81	0.59	50.79	0.55	0.60	SANet	42.36	0.9881	36.56	0.9583	33.53	0.9419	DKN	1.62	3.26	6.51
Dif-Pan	**43.89**	**1.03**	**0.83**	**50.22**	**0.72**	**0.53**	**52.33**	**0.46**	**0.51**	DuDoNet	44.96	0.9919	38.99	0.9744	35.43	0.9533	SGNet	**1.10**	2.44	4.77
MAG-Net	43.76	**1.03**	**0.83**	49.89	0.75	0.54	51.95	0.47	0.55	MAG-Net–	**45.03**	**0.9934**	**39.02**	**0.9780**	**35.62**	**0.9580**	MAG-Net	1.21	**2.40**	**4.39**

TABLE III
ABLATION STUDY ON THE TWO PROPOSED MODULES. BEST RESULTS ARE IN **BOLD**; SECOND BEST ARE UNDERLINED.

Module		Pansharpening			MR Image SR			Depth Image SR		
MPGM	MGRM	WV4	QB	GF1	2×	4×	8×	4×	8×	16×
✗	✗	43.68	49.94	51.60	44.32	38.92	35.29	1.50	2.79	5.12
✓	✗	43.83	50.09	52.34	44.46	39.16	35.35	1.41	2.63	4.75
✗	✓	43.71	50.00	51.71	44.13	38.96	35.31	1.47	2.69	4.88
✓	✓	**44.01**	**50.15**	**52.52**	**44.64**	**39.21**	**35.60**	**1.29**	**2.47**	**4.45**

TABLE IV
ABLATION STUDY ON THE GISR TASK COMBINATIONS. P, M, AND D REFER TO PANSHARPENING, MR IMAGE SR, AND DEPTH IMAGE SR, RESPECTIVELY. BEST RESULTS ARE IN **BOLD**; SECOND BEST ARE UNDERLINED.

Task			Pansharpening			MR Image SR			Depth Image SR		
P	M	D	WV4	QB	GF1	2×	4×	8×	4×	8×	16×
✓	✗	✗	43.73	49.89	51.98	–	–	–	–	–	–
✗	✓	✗	–	–	–	**45.03**	39.02	35.62	–	–	–
✗	✗	✓	–	–	–	–	–	–	**1.21**	**2.40**	**4.39**
✓	✓	✗	44.01	50.00	52.41	44.88	39.25	35.66	–	–	–
✓	✗	✓	**44.09**	50.05	**52.59**	–	–	–	1.28	2.48	4.50
✗	✓	✓	–	–	–	44.83	**39.28**	**35.70**	1.27	2.47	4.48
✓	✓	✓	44.01	**50.15**	52.52	44.64	39.21	35.60	1.29	2.47	4.45

TABLE V
ABLATION STUDY ON THE DATASET DIVERSITY. BEST RESULTS ARE IN **BOLD**; SECOND BEST ARE UNDERLINED.

Number	Pansharpening			MR Image SR			Depth Image SR		
	WV4	QB	GF1	2×	4×	8×	4×	8×	16×
3	43.75	–	–	–	39.05	–	1.35	–	–
6	43.87	49.86	–	44.49	39.03	–	1.35	2.61	–
9	**44.01**	**50.15**	**52.52**	**44.64**	**39.21**	**35.60**	**1.29**	**2.47**	**4.45**

TABLE VI
ABLATION STUDY ON THE GUIDED MODALITY OF PROMPT GENERATION. BEST RESULTS ARE IN **BOLD**; SECOND BEST ARE UNDERLINED.

Modality		Pansharpening			MR Image SR			Depth Image SR		
Text	Image	WV4	QB	GF1	2×	4×	8×	4×	8×	16×
✗	✗	43.81	50.01	50.73	44.19	39.02	35.56	1.47	2.72	4.91
✓	✗	43.74	**50.17**	52.31	44.10	38.92	35.57	1.46	2.75	4.84
✗	✓	43.73	50.09	52.43	44.22	39.04	**35.61**	1.48	2.68	4.91
✓	✓	**44.01**	50.15	**52.52**	**44.64**	**39.21**	35.60	**1.29**	**2.47**	**4.45**

imbalances where some tasks are well-optimized while others suffer from suboptimal restoration quality. In contrast, when both image and text modalities are jointly utilized for prompt generation, the model is able to construct more discriminative and task-aware prompts. Consequently, the SR performance of all tasks is improved in a more balanced and consistent manner.

e) The position of the routing module: Based on the observations in Figure 2, we found that the feature independence across different GISR subtasks is more pronounced in the encoder than in the decoder. This suggests that task-specific characteristics are more effectively distinguishable at earlier stages of the network. Motivated by this insight, we strategically place our routing module, MGPM, within the encoder to better exploit this inherent feature separation and perform task-adaptive modulation at an early stage.

To validate the effectiveness of this design, we compare two model variants: one with MGPM applied only in the decoder, one with MGPM applied in both the encoder and decoder. The results are presented in Table VII. As shown, placing MGPM in the encoder alone leads to noticeably better SR performance compared to applying it solely in the decoder, confirming the advantage of early-stage feature routing. Although incorporating MGPM into both the encoder and decoder yields marginally higher performance, the improvement is relatively minor and comes at the cost of increased model complexity and parameter count. Given the trade-off between accuracy and efficiency, we ultimately adopt the configuration where MGPM is placed only in the encoder.

V. CONCLUSION AND LIMITATION

In this paper, we propose MAG-Net, the first all-in-one Guided Image Super-Resolution (GISR) model that unifies multiple GISR subtasks within a single, coherent framework. To achieve task-specific adaptability and cross-modal generalization, we introduce two key modules: the *Multi-modal Prompt Generation Module* (MPGM), which dynamically generates task-aware prompts from the input image pair and predefined textual descriptions, and the *Multi-Guidance Routing Module* (MGRM), which leverages Mixture-of-Experts (MoE) techniques to adaptively route features based on the provided guidance. Extensive experiments across various GISR benchmarks validate the effectiveness of these modules and demonstrate that MAG-Net consistently outperforms existing

TABLE VII
ABLATION STUDY ON THE ROUTING POSITION. BEST RESULTS ARE IN **BOLD**; SECOND BEST ARE UNDERLINED.

Position	Pansharpening			MR Image SR			Depth Image SR		
	WV4	QB	GF1	2×	4×	8×	4×	8×	16×
Encoder	44.01	50.15	52.52	44.64	39.21	35.60	**1.29**	**2.47**	**4.45**
Decoder	43.53	49.97	51.93	44.20	39.01	35.59	1.50	2.74	4.94
Both	**44.03**	**50.16**	**52.57**	**44.65**	**39.23**	**35.63**	**1.29**	2.48	4.47

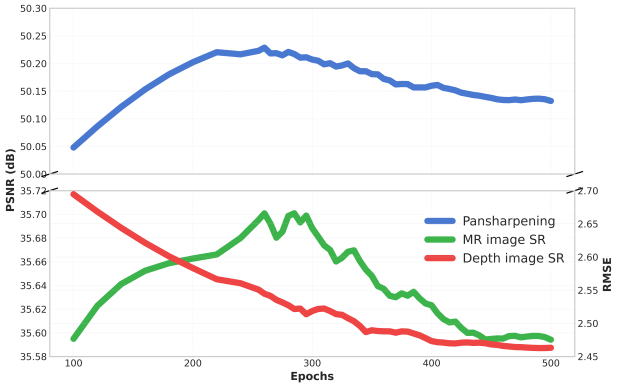

Fig. 11. Performance trends of different tasks across training epochs. Pansharpening and MR image SR are assessed using PSNR (higher values indicate better performance), whereas depth image SR is evaluated based on RMSE (lower values indicate better performance).

approaches under both all-in-one and subtask-specific training settings.

Despite the promising results, our method still faces certain challenges. As shown in Figure 11, during training, particularly in the later stages, the optimization process tends to favor the depth image SR task, leading to performance degradation in the other subtasks. This imbalance highlights a common issue in multi-task learning, where dominant tasks may overshadow others during joint optimization. In future work, we plan to explore more advanced task balancing strategies, such as dynamic loss weighting or gradient normalization techniques, to further enhance the fairness and robustness of the unified GISR framework.

REFERENCES

[1] S. Peng, X. Zhu, H. Deng, Z. Lei, and L.-J. Deng, "Fusionmamba: Efficient image fusion with state space model," *CoRR*, 2024.

[2] Z. Cao, S. Cao, L.-J. Deng, X. Wu, J. Hou, and G. Vivone, "Diffusion model with disentangled modulations for sharpening multispectral and hyperspectral images," *Information Fusion*, vol. 104, p. 102158, 2024.

[3] G. Yang, L. Zhang, A. Liu, X. Fu, X. Chen, and R. Wang, "Mgdun: An interpretable network for multi-contrast mri image super-resolution reconstruction," *Computers in Biology and Medicine*, vol. 167, p. 107605, 2023.

[4] P. Lei, L. Hu, F. Fang, and G. Zhang, "Joint under-sampling pattern and dual-domain reconstruction for accelerating multi-contrast mri," *IEEE Transactions on Image Processing*, 2024.

[5] Z. Wang, Z. Yan, and J. Yang, "Sgnet: Structure guided network via gradient-frequency awareness for depth map super-resolution," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 5823–5831.

[6] H. Lu, Y. Yang, S. Huang, X. Chen, B. Chi, A. Liu, and W. Tu, "Awfln: An adaptive weighted feature learning network for pansharpening," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[7] G. Li, J. Lv, Y. Tian, Q. Dou, C. Wang, C. Xu, and J. Qin, "Transformer-empowered multi-scale contextual matching and aggregation for multi-contrast mri super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 636–20 645.

[8] M. Zhou, K. Yan, J. Pan, W. Ren, Q. Xie, and X. Cao, "Memory-augmented deep unfolding network for guided image super-resolution," *International Journal of Computer Vision*, vol. 131, no. 1, pp. 215–242, 2023.

[9] B. Li, X. Liu, P. Hu, Z. Wu, J. Lv, and X. Peng, "All-in-one image restoration for unknown corruption," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 452–17 462.

[10] T. Wang, K. Zhang, Z. Shao, W. Luo, B. Stenger, T. Lu, T.-K. Kim, W. Liu, and H. Li, "Gridformer: Residual dense transformer with grid structure for image restoration in adverse weather conditions," *International Journal of Computer Vision*, pp. 1–23, 2024.

[11] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proceedings of IEEE International Conference on Computer Vision*. IEEE, 1998, pp. 839–846.

[12] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Transactions on Graphics*, vol. 26, no. 3, pp. 96–es, 2007.

[13] X. Shen, Q. Yan, L. Xu, L. Ma, and J. Jia, "Multispectral joint image restoration via optimizing a scale map," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 12, pp. 2518–2530, 2015.

[14] P. Song, X. Deng, J. F. Mota, N. Deligiannis, P. L. Dragotti, and M. R. Rodrigues, "Multimodal image super-resolution via joint sparse representations induced by coupled dictionaries," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 57–72, 2019.

[15] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, "Pannet: A deep network architecture for pan-sharpening," in *Proceedings of IEEE International Conference on Computer Vision*, 2017, pp. 5449–5457.

[16] Y. Wang, L.-J. Deng, T.-J. Zhang, and X. Wu, "SSconv: Explicit spectral-to-spatial convolution for pansharpening," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4472–4480.

[17] I. Marivani, E. Tsiligianni, B. Cornelis, and N. Deligiannis, "Multimodal deep unfolding for guided image super-resolution," *IEEE Transactions on Image Processing*, vol. 29, pp. 8443–8456, 2020.

[18] X. Deng and P. L. Dragotti, "Deep coupled ISTA network for multi-modal image super-resolution," *IEEE Transactions on Image Processing*, vol. 29, pp. 1683–1698, 2019.

[19] G. Yang, M. Zhou, K. Yan, A. Liu, X. Fu, and F. Wang, "Memory-augmented deep conditional unfolding network for pan-sharpening," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1788–1797.

[20] P. Lei, F. Fang, G. Zhang, and M. Xu, "Deep unfolding convolutional dictionary model for multi-contrast MRI super-resolution and reconstruction," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 8 2023, pp. 1008–1016.

[21] J. Zhang, J. Huang, M. Yao, Z. Yang, H. Yu, M. Zhou, and F. Zhao, "Ingredient-oriented multi-degradation learning for image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5825–5835.

[22] V. Potlapalli, S. W. Zamir, S. H. Khan, and F. Shahbaz Khan, "Promptir: Prompting for all-in-one image restoration," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

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

[24] X. Chen, Y. Liu, Y. Pu, W. Zhang, J. Zhou, Y. Qiao, and C. Dong, "Learning a low-level vision generalist via visual task prompt," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 2671–2680.

[25] Y. Ai, H. Huang, X. Zhou, J. Wang, and R. He, "Multimodal prompt perceiver: Empower adaptiveness generalizability and fidelity for all-in-one image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 25 432–25 444.

[26] S. Cao, Y. Liu, W. Zhang, Y. Qiao, and C. Dong, "Grids: Grouped multiple-degradation restoration with image degradation similarity," in *European Conference on Computer Vision*. Springer, 2025, pp. 70–87.

[27] Z. Yang, H. Chen, Z. Qian, Y. Yi, H. Zhang, D. Zhao, B. Wei, and Y. Xu, "All-in-one medical image restoration via task-adaptive routing," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 67–77.

[28] C. Ma, Z. Li, J. He, J. Zhang, Y. Zhang, and H. Shan, "Prompted contextual transformer for incomplete-view ct reconstruction," *arXiv preprint arXiv:2312.07846*, 2023.

[29] Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjölund, and T. B. Schön, "Controlling vision-language models for universal image restoration," *arXiv preprint arXiv:2310.01018*, vol. 3, no. 8, 2023.
[30] X. Zhang, J. Ma, G. Wang, Q. Zhang, H. Zhang, and L. Zhang, "Perceive-ir: Learning to perceive degradation better for all-in-one image restoration," *IEEE Transactions on Image Processing*, 2025.
[31] Q. Yan, A. Jiang, K. Chen, L. Peng, Q. Yi, and C. Zhang, "Textual prompt guided image restoration," *Engineering Applications of Artificial Intelligence*, vol. 155, p. 110981, 2025.
[32] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5728–5739.
[33] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.
[34] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *arXiv preprint arXiv:1701.06538*, 2017.
[35] X. Meng, Y. Xiong, F. Shao, H. Shen, W. Sun, G. Yang, Q. Yuan, R. Fu, and H. Zhang, "A large-scale benchmark data set for evaluating pansharpening performance: Overview and implementation," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 1, pp. 18–52, 2020.
[36] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European Conference on Computer Vision*. Springer, 2012, pp. 746–760.
[37] B. Kim, J. Ponce, and B. Ham, "Deformable kernel networks for joint image filtering," *International Journal of Computer Vision*, vol. 129, no. 2, pp. 579–600, 2021.
[38] J. M. J. Valanarasu, R. Yasarla, and V. M. Patel, "Transweather: Transformer-based restoration of images degraded by adverse weather conditions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2353–2363.
[39] H. Gao, J. Yang, Y. Zhang, N. Wang, J. Yang, and D. Dang, "Prompt-based ingredient-oriented all-in-one image restoration," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.
[40] Y. Cui, S. W. Zamir, S. Khan, A. Knoll, M. Shah, and F. S. Khan, "Adair: Adaptive all-in-one image restoration via frequency mining and modulation," *arXiv preprint arXiv:2403.14614*, 2024.
[41] Q. Jia, X. Wan, B. Hei, and S. Li, "Dispnet based stereo matching for planetary scene depth estimation using remote sensing images," in *2018 10th IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS)*. IEEE, 2018, pp. 1–5.
[42] Z.-R. Jin, T.-J. Zhang, T.-X. Jiang, G. Vivone, and L.-J. Deng, "Lagconv: Local-context adaptive convolution kernels with global harmonic bias for pansharpening," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 1113–1121.
[43] Z. Li, H. Huang, Y. Li, and Y. Pan, "M3dnet: A manifold-based discriminant feature learning network for hyperspectral imagery," *Expert Systems with Applications*, vol. 144, p. 113089, 2020.
[44] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9413–9422.
[45] L. Lu, W. Li, X. Tao, J. Lu, and J. Jia, "Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6368–6377.
[46] C.-M. Feng, Y. Yan, K. Yu, Y. Xu, H. Fu, J. Yang, and L. Shao, "Exploring separable attention for multi-contrast mr image super-resolution," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.
[47] X. Wang, X. Chen, B. Ni, Z. Tong, and H. Wang, "Learning continuous depth representation via geometric spatial aggregator," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2023, pp. 2698–2706.
[48] Z. Zhong, X. Liu, J. Jiang, D. Zhao, and X. Ji, "Deep attentional guided image filtering," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
[49] Z. Zhong, X. Liu, J. Jiang, D. Zhao, Z. Chen, and X. Ji, "High-resolution depth maps imaging via attention-based hierarchical multi-modal fusion," *IEEE Transactions on Image Processing*, vol. 31, pp. 648–663, 2021.
[50] Z. Zhao, J. Zhang, S. Xu, Z. Lin, and H. Pfister, "Discrete cosine transform network for guided depth map super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5697–5707.