

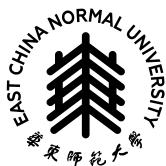
2026 届研究生硕士学位论文

分 类 号: \_\_\_\_\_

学校代码: 10269

密 级: \_\_\_\_\_

学 号: 51265901080



华东師範大學

East China Normal University

硕士学位论文

MASTER'S DISSERTATION

论文题目： 基于动态路由与多模态提示  
的一体化引导图像超分辨率  
研究

院 系: 计算机科学与技术学院

专 业: 计算机技术

研 究 方 向: 图像处理

学位申请 人: 王君

指 导 教 师: 方发明 教授

2026 年 02 月 12 日

Dissertation for Master's Degree in 2026

University Code: 10269

Student ID: 51265901080

## EAST CHINA NORMAL UNIVERSITY

**Title: Research on All-in-One Guided Image  
Super-Resolution Based on Dynamic  
Routing and Multi-modal Prompting**

---

Department / School: School of Computer Science and Technology

Major: Computer Science

Research Direction: Image Processing

Candidate: Jun Wang

Supervisor: Prof. Faming Fang

## 华东师范大学学位论文原创性声明

郑重声明：本人呈交的学位论文《基于动态路由与多模态提示的一体化引导图像超分辨率研究》，是在华东师范大学攻读硕士/博士（请勾选）学位期间，在导师的指导下进行的研究工作及取得的研究成果。除文中已经注明引用的内容外，本论文不包含其他个人已经发表或撰写过的研究成果。对本文的研究做出重要贡献的个人和集体，均已在文中作了明确说明并表示谢意。

作者签名： 日 期： 年 月 日

## 华东师范大学学位论文著作权使用声明

《基于动态路由与多模态提示的一体化引导图像超分辨率研究》系本人在华东师范大学攻读学位期间在导师指导下完成的硕士/博士（请勾选）学位论文，本论文的著作权归本人所有。本人同意华东师范大学根据相关规定保留和使用此学位论文，并向主管部门和学校指定的相关机构送交学位论文的印刷版和电子版；允许学位论文进入华东师范大学图书馆及数据库被查阅、借阅；同意学校将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于（请勾选）

- ( ) 1. 经华东师范大学相关部门审查核定的“内部”或“涉密”学位论文 \*，  
于 年 月 日解密，解密后适用上述授权。  
( ) 2. 不保密，适用上述授权。

导师签名： 作者签名：

日 期： 年 月 日

\* “涉密”学位论文应是已经华东师范大学学位管理办公室或保密委员会审定过的学位论文（需附获批的《华东师范大学研究生申请学位论文“涉密”审批表》方为有效），未经上述部门审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权）。

## 王君 硕士学位论文答辩委员会成员名单

姓名	职称	单位	备注
赵六	教授	华东师范大学	主席

## 摘 要

引导图像超分辨率旨在利用高分辨率的辅助图像（如全色图像、RGB 图像等）来提升低分辨率目标图像的空间分辨率，被广泛应用于遥感监测、医学成像及深度估计等领域。然而，现有的引导超分辨率方法主要针对单一特定任务设计，由于不同任务间存在巨大的模态差异和成像机理鸿沟，导致模型在跨任务场景下泛化能力不足，且面临着任务间相互干扰的挑战。为打破“一任务一模型”的传统范式，实现多模态任务的高效协同与通用重建，本文基于提示学习与混合专家机制，提出了层层递进的两种一体化引导图像超分辨率方法，具体如下：

(1) 提出一种基于视觉特征引导的动态路由重建方法。针对现有单一网络在处理多任务时易产生特征冲突与负迁移，且未能充分解耦不同任务特性的问题，本文提出了一种基于视觉感知的动态路由机制。该方法引入混合专家架构 (MoE)，构建多引导路由模块，利用图像自身的视觉特征作为隐式引导信号，自适应地激活适合当前输入的专家网络路径。该方法在不显著增加计算成本的前提下，实现了对不同任务特征的差异化处理，有效缓解了多任务学习中的干扰问题，为一体化模型的构建奠定了结构基础。

(2) 提出一种融合文本语义的多模态提示驱动重建方法。针对仅依赖视觉特征进行引导时对任务意图理解不足，且难以应对复杂模态差异的瓶颈，本文在动态路由的基础上，引入文本语义先验，构建了多模态提示生成模块。该方法创新性地将任务描述文本 (Textual Description) 映射至语义空间，并与视觉特征深度融合，生成显式的多模态任务指令 (Prompts)。这些指令如同“导航员”一般，精准调控网络内部的特征流向与交互方式。实验表明，该方法实现了领域视觉信息与高层语义知识的深度融合，显著提升了模型在全色锐化、深度图超分及磁共振重建等多个任务上的性能与泛化能力。

(3) 设计并实现了一体化引导图像超分辨率算法验证与可视化系统。针对现有理论研究缺乏统一的评估平台，且难以直观展示模型内部动态机制与多任务处理效果的问题，本文基于所提出的算法模型，研发了一个集算法验证、对比分析与可视化展示于一体的实验系统。该系统完整集成了本文提出的两种核心算法，支

持多源异构数据的统一接入与一键处理，并特别设计了中间特征（如动态路由分布、多模态提示热力图）的可视化模块。通过该系统，不仅直观验证了所提算法在实际应用场景下的有效性与鲁棒性，也增强了深度模型的透明度与可解释性，为相关技术的工程化应用提供了有力的工具支撑。

**关键词：**引导图像超分辨率，一体化模型，动态路由，多模态融合

## Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequo doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et.

**Keywords:** *To, be, or, not, to, be*

## 目 录

摘要 .....	1
Abstract .....	3
插图目录 .....	6
表格目录 .....	7
符号表 .....	8
第一章 绪 论 .....	1
第二章 相关知识及研究现状 .....	2
2.1 正文子标题 .....	2
2.1.1 正文子子标题 .....	2
第三章 面向一体化任务的视觉感知与动态路由 GISR 方法 .....	3
3.1 引言 .....	3
3.2 方法 .....	5
3.2.1 整体网络架构 .....	5
3.2.2 基于视觉感知的动态路由模块 .....	8
3.2.3 损失函数 .....	10
3.3 实验设置 .....	11
3.3.1 数据集 .....	11
3.3.2 评测指标 .....	13
3.3.3 实验细节 .....	13
3.4 实验结果 .....	14
3.4.1 对比实验 .....	14
All-in-One 方法对比 .....	14
One-by-One 方法对比 .....	16
3.4.2 消融实验 .....	18
VPRM 有效性消融分析 .....	18
VPRM 位置消融分析 .....	18
3.4.3 动态路由机制的可视化分析 .....	19
3.5 本章小结 .....	20
第四章 融合多模态语义提示的一体化 GISR 方法 .....	22

4.1 引言 .....	22
4.2 方法 .....	23
4.2.1 整体网络架构 .....	24
4.2.2 多模态提示生成模块 .....	25
4.2.3 融合多模态提示的动态路由模块 .....	27
4.3 实验设置 .....	29
4.4 实验结果 .....	29
4.4.1 .....	29
4.5 本章小结 .....	29
<b>第五章 一体化引导图像超分辨率系统</b> .....	<b>30</b>
5.1 引言 .....	30
5.2 需求分析 .....	30
5.3 系统设计 .....	30
5.4 开发环境和依赖 .....	30
5.5 本章小结 .....	30
<b>第六章 总结与展望</b> .....	<b>31</b>
6.1 工作总结 .....	31
6.2 未来展望 .....	31
<b>参考文献</b> .....	<b>32</b>
<b>致谢</b> .....	<b>33</b>
<b>附录</b> .....	<b>34</b>
7.1 附录子标题 .....	34
7.1.1 附录子子标题 .....	34
<b>攻读硕/博士学位期间科研情况</b> .....	<b>35</b>

## 插图目录

图 3.1 One-by-One 单任务模型与 All-in-One 一体化多任务模型的范式对比 . . . . .	3
图 3.2 使用编码器-解码器网络架构处理不同的 GISR 子任务时，某些通道的特征图的 L2 范数分布 . . . . .	4
图 3.3 VPNet 网络整体架构 . . . . .	5
图 3.4 视觉感知路由模块（VPRM）结构示意图 . . . . .	8
图 3.5 三种 GISR 子任务的数据集示例。每一行展示一个任务（从上至下依次为：全色锐化、磁共振图像超分、深度图超分），包括低分辨率输入、高分辨率引导图像及高分辨率真值。 . . . . .	12
图 3.6 动态路由机制激活热力图 . . . . .	20
图 4.1 纯视觉感知面临的歧义性挑战与多模态语义提示的引入 . . . . .	23
图 4.2 MAG-Net 网络整体架构示意图。模型接收低分辨率图像、高分辨率引导图像及任务语义描述作为输入，通过多模态提示生成模块（MPGM）提取语义先验，并结合视觉特征在 MGRM 中动态激活特定的专家网络 . . . . .	24
图 4.3 多模态提示生成模块（MPGM）处理流程示意图 . . . . .	26
图 4.4 融合多模态提示的动态路由模块（MGRM）结构示意图 . . . . .	28
图 7.1 图片测试 . . . . .	34

## 表格目录

表 3.1 不同 all-in-one 方法在 Pansharpening 任务上的对比结果, 最优结果用 粗体表示, 次优结果用 <u>下划线</u> 表示 .....	15
表 3.2 不同 all-in-one 方法在 MR Image SR 任务上的对比结果, 最优结果用 粗体表示, 次优结果用 <u>下划线</u> 表示 .....	15
表 3.3 不同 all-in-one 方法在 Depth Image SR 任务上的对比结果, 最优结果用 粗体表示, 次优结果用 <u>下划线</u> 表示 .....	16
表 3.4 不同 one-by-one 方法在 Pansharpening 任务上的对比结果, 最优结果用 粗体表示, 次优结果用 <u>下划线</u> 表示 .....	17
表 3.5 不同 one-by-one 方法在 MR Image SR 任务上的对比结果, 最优结果用 粗体表示, 次优结果用 <u>下划线</u> 表示 .....	17
表 3.6 不同 one-by-one 方法在 Depth Image SR 任务上的对比结果, 最优结果用 粗体表示, 次优结果用 <u>下划线</u> 表示 .....	17
表 3.7 VPRM 模块有效性消融实验结果分析 .....	18
表 3.8 不同模块位置 (Encoder/Decoder) 对各任务性能的影响分析 .....	19

## 符号表

DFT	密度泛函理论 (Density functional theory)
DMRG	密度矩阵重正化群密度矩阵重正化群密度矩阵重正化群 (Density-Matrix Reformation-Group)
RAII	资源获取即初始化 (Resource Acquisition Is Initialization)

# 第一章 绪 论

## 第二章 相关知识及研究现状

### 2.1 正文子标题

#### 2.1.1 正文子子标题

- 1) 自定义列表编号与缩进
- 2) 自定义列表编号与缩进

### 第三章 面向一体化任务的视觉感知与动态路由 GISR 方法

#### 3.1 引言

引导图像超分辨率 (Guided Image Super-Resolution, GISR) 旨在利用高分辨率的辅助模态图像 (如全色图像、RGB 图像等) 来指导低分辨率目标图像的重建, 在遥感对地观测、医学影像分析以及三维深度估计等领域发挥着至关重要的作用。随着深度学习技术的发展, 针对单一特定任务 (如全色锐化、磁共振图像超分或深度图超分) 的专用模型已经取得了显著的性能提升。然而, 在实际应用场景中, 往往需要处理来自不同传感器、不同模态以及不同放大倍率的多源数据。如图 3.1 (a) 所示, 传统的“一任务一模型”范式不仅导致了巨大的存储与计算开销, 也难以挖掘不同 GISR 子任务之间潜在的共性特征。因此, 如图 3.1 (b) 所示, 构建一个能够同时处理多种 GISR 任务的一体化 (All-in-One) 模型, 已成为当前图像复原领域的重要研究趋势。

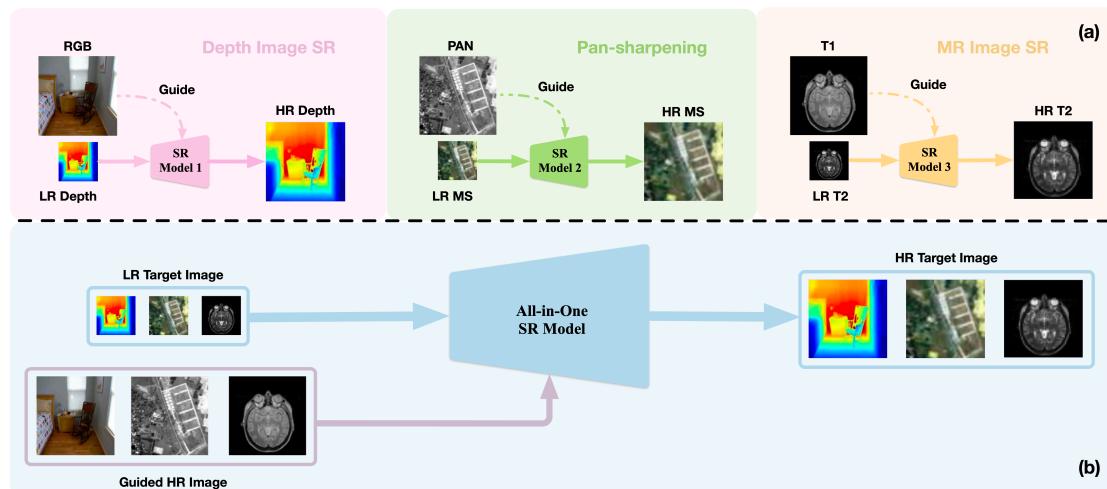


图 3.1 One-by-One 单任务模型与 All-in-One 一体化多任务模型的范式对比

尽管一体化模型具有极高的应用价值, 但其训练过程面临着严峻的挑战。与去噪、去雨等自然图像复原任务不同, GISR 涵盖了全色影像、深度图和磁共振影像等多种差异巨大的数据模态。这些模态在成像机理、频谱特性以及纹理结构上存在显著的分布差异 (Domain Gap)。如果简单地将所有任务的数据混合, 强制使用一个共享参数的骨干网络 (Shared Backbone) 进行训练, 不同任务的梯度更新方向往往会发生冲突, 导致模型参数在优化过程中产生震荡。这种现象被称为

“任务干扰”或“负迁移”，不同任务的优化方向不同，导致一体化模型在某些子任务上的性能反而低于单独训练的专用模型。

在设计一体化网络之前，我们需要先深入理解不同 GISR 子任务在特征空间中的分布特性。为了探究多任务混合训练时潜在的干扰机制，我们对一个本章方法所使用的基础 Encoder-Decoder 网络模型 Restormer，在处理不同任务时的中间特征进行了统计分析。

图 3.2 展示了不同任务在编码-解码架构中某些通道的  $L_2$  范数分布情况。从图中可以清晰地观察到，不同任务（如全色锐化、磁共振超分）在同一通道上的激活强度存在显著差异。例如，在第 8 通道，全色锐化任务表现出极高的激活值，而深度图超分任务的激活值则接近于零。

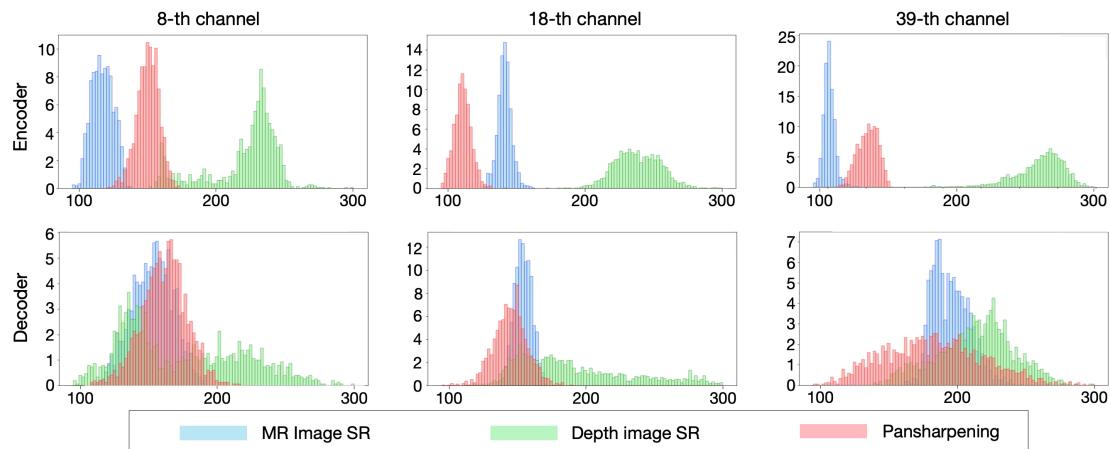


图 3.2 使用编码器-解码器网络架构处理不同的 GISR 子任务时，某些通道的特征图的  $L_2$  范数分布

为了解决上述“任务干扰”问题，打破多任务协同训练的瓶颈，并充分利用高分辨率引导图片的指导能力，本章提出了一种面向一体化任务的视觉感知与动态路由 GISR 方法 (VPNet)。该方法摒弃了传统静态网络的参数共享机制，引入了混合专家 (Mixture-of-Experts, MoE) 架构思想，旨在通过网络结构的动态化来实现特征的物理解耦。具体而言，我们设计了一个视觉感知路由模块 (Visual-Perception Routing Module, VPRM)，该模块能够充当“交通指挥员”的角色。它利用输入图像自身的底层视觉特征 (如纹理、边缘、灰度分布等)，以及引导图片中提取的视觉特征，共同作为隐式的感知信号，为每个像素动态分配最匹配的专家网络 (Experts) 路径。

通过这种基于视觉感知的动态路由机制，模型能够根据输入数据的模态特性，自适应地激活特定的计算分支。例如，全色锐化任务倾向于激活处理高频细节的专家，而深度图超分任务则可能更多地依赖处理平滑梯度的专家。这种机制在不显著增加计算量的前提下，有效地隔离了不同任务间的特征干扰，使得模型既能保留对共性特征的学习能力，又能兼顾不同模态的特异性需求。本章将详细阐述该方法的网络架构设计、动态路由机制的实现细节，并通过一系列实验验证其在一体化 GISR 任务中的有效性。

## 3.2 方法

本章的研究重点在于利用底层视觉特征引导一体化模型高效处理多模态任务，从而缓解不同 GISR 子任务间的参数干扰问题。为实现该目标，提出了一种面向一体化任务的视觉感知与动态路由 GISR 方法 (VP-Net)，其整体架构如图 3.1 所示。该方法摒弃了传统的静态共享参数模式，通过引入视觉感知路由机制，根据输入图像的纹理和结构特性动态规划特征提取路径，实现任务特征的物理解耦。本章将在小节 3.2.1 节首先阐述模型的整体网络架构；随后在小节 3.2.2 节详细介绍核心组件——基于视觉感知的动态路由模块的设计细节与工作机理；最后在小节 3.2.3 节给出用于指导模型优化的损失函数定义。

### 3.2.1 整体网络架构

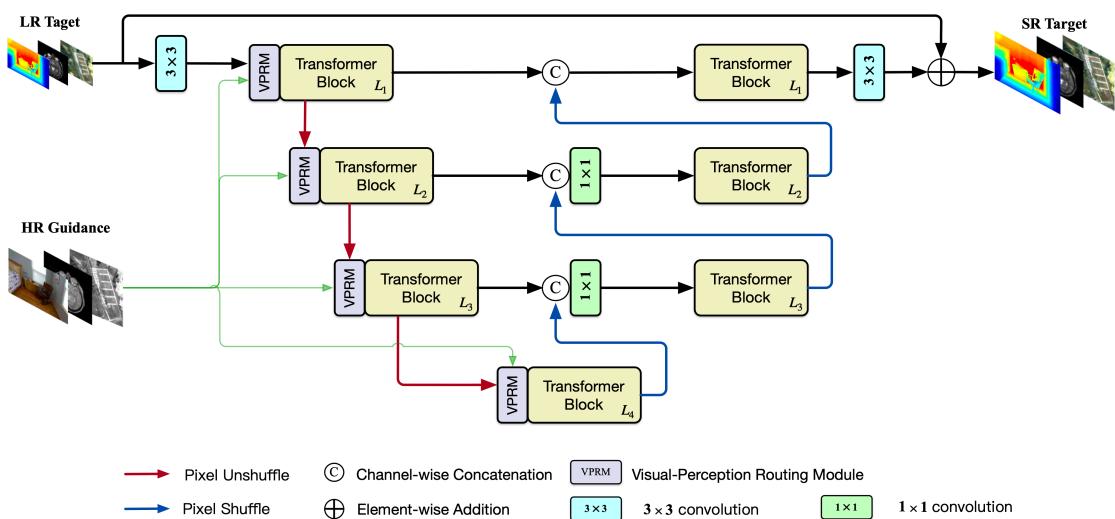


图 3.3 VPNet 网络整体架构

如图 3.3 所示，本章提出的 VP-Net 建立在一个分层的 U 型编码器-解码器（Encoder-Decoder）架构之上。该骨干网络设计灵感来源于 Restormer，旨在利用其高效的 Transformer 模块处理高分辨率特征，并通过多尺度的特征交互来捕捉图像的长距离依赖与局部细节。VP-Net 主要由三个部分组成：浅层特征提取模块、嵌入视觉感知路由的编码器、以及图像重建解码器。

模型的输入包含两个部分：待恢复的低分辨率目标图像  $I_{\text{LR}} \in \mathbb{R}^{H \times W \times C_{\text{in}}}$  和高分辨率引导图像  $I_{\text{HR}} \in \mathbb{R}^{sH \times sW \times C_{\text{guide}}}$ （其中  $s$  为超分倍率）。

由于  $I_{\text{LR}}$  和  $I_{\text{HR}}$  来自不同的模态（如全色、深度或磁共振），且具有不同的空间分辨率，我们首先使用两个独立的  $3 \times 3$  卷积层作为浅层特征提取器，分别将它们映射到统一的通道维度  $C$ 。对于  $I_{\text{LR}}$ ，提取得到的特征  $F_0$  将作为主干网络的初始输入；对于  $I_{\text{HR}}$ ，提取得到的特征  $F_{\text{guide}}$  将作为视觉引导信号，被送入后续的视觉感知路由模块中，用于指引特征的分流和 MoE 中专家的选择。

编码器旨在逐步降低特征的空间分辨率并增加通道数，以获取更大的感受野。它包含 4 个层级，每个层级由若干个 Transformer Block 堆叠而成。这些 Block 采用了 Restormer 的核心组件——多头转置注意力机制（MDTA）和门控前馈网络（GDFN），以在特征维度上高效聚合上下文信息。与传统共享参数的 Restormer 不同，为了解决一体化训练中的任务干扰问题，我们在编码器的每个层级中嵌入了视觉感知路由模块（Visual-Aware Routing Module, VARM）。

在网络结构中，VARM 被放置在每个层级的 Transformer Block 之后。该模块同时接收当前层级输出的主干特征  $F_l^{\text{enc}}$  与经过下采样的引导特征  $F_{\text{guide}}^l$ ，其中  $l$  表示层级索引。 $F_{\text{guide}}$  中蕴含的丰富纹理、边缘和结构信息在此充当“视觉罗盘”的角色，VARM 通过融合主干特征与引导特征的统计信息，动态地计算像素级的门控权重，从而决定每个空间位置的特征应被分配至哪一组专家网络进行处理。通过这种视觉感知驱动的动态路由，网络能够依据输入图像的底层视觉特性，将特征动态分流至最适合当前纹理分布的专家网络（Experts）中。这种机制在编码阶段实现了物理路径的解耦，使得不同任务能够激活不同的参数子集。例如，全色锐化任务所涉及的高频空间细节特征会被路由至擅长处理精细纹理的专家分支，而深度图超分任务中的平滑梯度特征则会被导向另一组专家。由此，不同任务的梯度更新在参数空间中被有效隔离，从根本上缓解了梯度冲突与负迁移问题。在编码器的各层级之间，空间下采样操作通过 Pixel-Unshuffle 实现。该操作将空间

维度为  $H \times W$  的特征图重组为  $\frac{H}{2} \times \frac{W}{2}$  的特征图，同时通道数增加为原来的 4 倍，随后通过一个  $1 \times 1$  卷积层将通道数调整至目标维度。相较于传统的步幅卷积（Strided Convolution）或池化（Pooling）操作，Pixel-Unshuffle 在减少空间维度的同时能够完整保留通道信息，有效避免了下采样过程中的信息丢失。

解码器由 4 个与编码器对称的层级组成，负责逐步恢复图像的空间分辨率。每个解码器层级包含对应数量的 Transformer Block，用于从深层编码特征中重建高频细节与精细纹理。为了弥补逐层下采样过程中不可避免的空间信息损失，网络引入了跳跃连接（Skip Connections），将编码器各层级的输出特征与解码器对应层级的输入特征在通道维度上进行拼接（Concatenation），并通过一个  $1 \times 1$  卷积层将拼接后的通道数压缩回当前层级的标准维度。跳跃连接使得解码器在重建过程中能够直接访问编码器提取的多尺度浅层特征，从而保留更丰富的空间结构信息，提升重建图像的质量。需要说明的是，基于前期的实验观察（详见本章消融实验部分），不同任务间的特征差异主要体现在特征提取与理解阶段，即编码器部分。在解码器阶段，来自不同任务的特征经过编码器的动态路由已经被充分解耦和提纯，此时对它们进行共享参数的重建处理并不会引发显著的干扰。因此，为了保持模型的计算效率，我们在解码器阶段并未引入路由模块，而是沿用了标准的静态 Transformer Block 结构。解码器各层级之间的特征上采样通过 Pixel-Shuffle 操作实现，该操作将通道维度上的冗余信息重组到空间维度，逐步将特征图放大至目标分辨率。

经过解码器处理后，网络得到与输入  $I_{LR}$  具有相同空间分辨率（即  $sH \times sW$ ）的深层特征表示。该特征通过一个  $3 \times 3$  卷积层映射回原始图像的通道空间，得到残差图像  $I_{res} \in \mathbb{R}^{sH \times sW \times C_{in}}$ 。最终的超分辨率重建结果  $I_{SR}$  通过残差学习策略获得，即将残差图像与经双线性插值上采样至目标分辨率的输入图像相加：

$$I_{SR} = I_{res} + \text{Upsample}(I_{LR}) \quad (3.1)$$

其中  $\text{Upsample}(\cdot)$  表示双线性插值上采样操作。这种残差学习范式使得网络只需学习低分辨率输入与高分辨率目标之间的高频差异部分，而非从零开始重建完整图像，显著降低了学习难度并加速了模型的收敛过程。

### 3.2.2 基于视觉感知的动态路由模块

视觉感知路由模块 (Visual-Perception Routing Module, VPRM) 是 VP-Net 实现一体化多任务处理的核心组件。该模块的设计目标是利用高分辨率引导图像中蕴含的纹理、边缘与结构信息作为“视觉罗盘”，为主干特征中的每个空间位置动态选择最合适专家网络进行处理，从而在参数空间中实现不同任务特征的物理隔离。VPRM 的结构如图 3.4 所示。

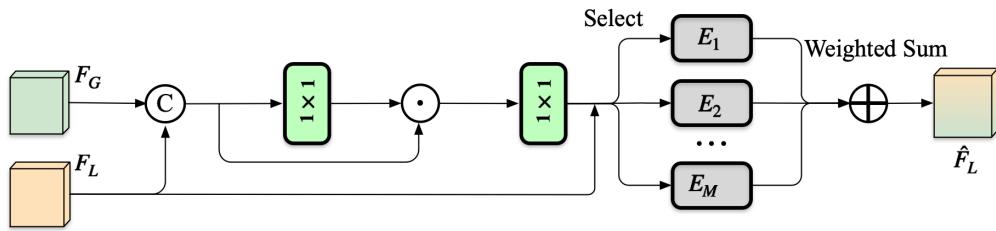


图 3.4 视觉感知路由模块 (VPRM) 结构示意图

VPRM 采用稀疏门控混合专家 (Sparsely-Gated Mixture-of-Experts) 架构，其内部包含三个关键组成部分：视觉感知门控网络 (Visual-Perception Gating Network)、专家网络组 (Expert Networks) 以及稀疏分发与加权聚合机制 (Sparse Dispatch & Weighted Aggregation)。

对于编码器第  $l$  层级，VPRM 接收两个输入：当前层级的主干特征  $F_L \in \mathbb{R}^{B \times C_l \times H_l \times W_l}$  以及经过同步下采样的引导特征  $F_G \in \mathbb{R}^{B \times C_l \times H_l \times W_l}$ 。视觉感知门控网络首先将两者在通道维度上进行拼接，得到融合特征：

$$F_{\text{fuse}} = \text{Conv}_{1 \times 1}([\text{Concat}(F_G, F_L)]) \in \mathbb{R}^{B \times 2C_l \times H_l \times W_l} \quad (3.2)$$

其中  $\text{Conv}_{1 \times 1}(\cdot)$  表示  $1 \times 1$  逐点卷积操作。该卷积层的作用在于对拼接后的引导特征与主干特征进行跨模态信息融合，使门控网络能够同时感知引导图像的高频结构信息与目标图像的当前特征状态。随后，融合特征  $F_{\text{fuse}}$  被重塑为像素级的特征向量序列  $\mathbf{f}_{\text{fuse}} \in \mathbb{R}^{N \times 2C_l}$  (其中  $N = B \times H_l \times W_l$  为总的空间像素数量)，并通过一个可学习的线性投影层  $\mathbf{W}_g \in \mathbb{R}^{2C_l \times M}$  (其中  $M$  为专家总数) 计算每个像素位置对各专家的路由分数：

$$\mathbf{s} = \mathbf{f}_{\text{fuse}} \mathbf{W}_g \in \mathbb{R}^{N \times M} \quad (3.3)$$

在训练阶段,为了增强路由决策的探索性并防止门控网络过早坍缩至固定的专家分配模式,VPRM引入了带噪声的Top- $k$ 门控机制(Noisy Top- $k$  Gating)。具体而言,通过另一组可学习参数 $\mathbf{W}_{\text{noise}} \in \mathbb{R}^{2C_l \times M}$ 生成依赖于输入的噪声标准差,并向路由分数中注入可控的高斯噪声:

$$\sigma = \text{Softplus}(\mathbf{f}_{\text{fuse}} \mathbf{W}_{\text{noise}}) + \varepsilon \quad (3.4)$$

$$\tilde{\mathbf{s}} = \mathbf{s} + \boldsymbol{\xi} \odot \sigma, \quad \boldsymbol{\xi} \sim \mathcal{N}(0, \mathbf{I}) \quad (3.5)$$

其中 $\varepsilon$ 为一个极小的正常数以保证数值稳定性, $\odot$ 表示逐元素乘法。基于带噪声的路由分数 $\tilde{\mathbf{s}}$ ,门控网络为每个像素选取得分最高的 $k$ 个专家(在本章实现中 $k=2$ ),并对这 $k$ 个专家的分数进行Softmax归一化以获得最终的门控权重:

$$G(\mathbf{f}_{\text{fuse}})_i = \begin{cases} \text{Softmax}(\tilde{\mathbf{s}}_i) & \text{if } i \in \text{TopK}(\tilde{\mathbf{s}}) \\ 0 & \text{otherwise} \end{cases} \quad (3.6)$$

其中TopK( $\cdot$ )返回分数最高的 $k$ 个专家的索引集合。未被选中的专家对应的门控权重为零,由此实现了稀疏激活——每个像素仅被路由至 $k$ 个专家,而非全部 $M$ 个专家,从而在引入多专家多样性的同时有效控制了计算开销。本章中,我们通过实验证证了 $k=2, M=4$ 的设置在性能与效率之间达到了较好的平衡,既能充分利用多专家的优势,又不会引入过多的计算负担。

专家网络组由 $M$ 个结构相同但参数独立的多层感知机(MLP)组成。每个专家 $E_i$ ( $i=1, 2, \dots, M$ )包含两层全连接层,中间以GELU激活函数连接:

$$E_i(\mathbf{x}) = \mathbf{W}_i^2 \cdot \text{GELU}(\mathbf{W}_i^1 \mathbf{x} + \mathbf{b}_i^1) + \mathbf{b}_i^2 \quad (3.7)$$

其中 $\mathbf{W}_i^1 \in \mathbb{R}^{C_l \times d_h}$ 、 $\mathbf{W}_i^2 \in \mathbb{R}^{d_h \times C_l}$ 为第 $i$ 个专家的权重矩阵, $d_h$ 为隐藏层维度(通过膨胀因子 $r$ 控制,即 $d_h = r \cdot C_l$ )。每个专家拥有独立的参数空间,能够学习特定的特征变换模式,例如某些专家可能专注于高频纹理的增强,而另一些专家则擅长处理低频平滑区域。

在稀疏分发阶段,VPRM根据门控权重将主干特征 $F_L$ 中的像素级特征向量分发至对应的专家网络。仅门控权重非零的像素-专家对会产生实际的计算,未被选中的专家不参与当前像素的处理。各专家独立地对分配给自身的特征子集进行变换后,其输出通过门控权重进行加权求和,得到最终的路由输出:

$$\hat{F}_L = \sum_{i=1}^M G(\mathbf{f}_{\text{fuse}})_i \cdot E_i(F_L) \quad (3.8)$$

最终, VPRM 的输出通过残差连接与原始主干特征相加, 确保信息流动的畅通性并降低训练难度:

$$F_L^{\text{out}} = \hat{F}_L + F_L \quad (3.9)$$

通过上述视觉感知驱动的动态路由机制, VPRM 能够根据引导图像所提供的视觉先验信息, 在像素级别对特征进行自适应的专家分配。不同模态、不同纹理特性的输入数据将被自动路由至不同的专家子集, 实现了特征提取路径的物理解耦, 从根本上缓解了一体化训练中的任务干扰与梯度冲突问题。

### 3.2.3 损失函数

VP-Net 的总体训练损失由两部分组成: 重建损失  $\mathcal{L}_1$  和负载均衡正则化损失  $\mathcal{L}_{\text{Balance}}$ , 其表达式为:

$$\mathcal{L} = \mathcal{L}_1 + \gamma \mathcal{L}_{\text{Balance}} \quad (3.10)$$

其中  $\mathcal{L}_1$  为像素级的  $L_1$  重建损失, 用于度量网络输出的超分辨率重建图像  $I_{\text{SR}}$  与对应的高分辨率真实标签  $I_{\text{GT}}$  之间的差异:

$$\mathcal{L}_1 = \| I_{\text{SR}} - I_{\text{GT}} \|_1 \quad (3.11)$$

$L_1$  损失相较于  $L_2$  损失对异常值更为鲁棒, 能够在保持整体重建精度的同时更好地恢复图像的高频纹理细节, 已被广泛应用于图像超分辨率任务中。

$\mathcal{L}_{\text{Balance}}$  是专门为混合专家 (MoE) 架构设计的负载均衡正则化项, 其目的在于防止门控网络在训练过程中出现“专家坍缩”现象——即大量像素被持续分配至少数几个专家, 而其余专家长期处于闲置状态, 导致模型的参数容量未能被充分利用。该损失由重要性损失 (Importance Loss) 和负载损失 (Load Loss) 两部分组成, 分别从权重分配和选择频率两个角度约束各专家的使用趋于均匀:

$$\mathcal{L}_{\text{Balance}} = \text{CV}^2 \left( \sum_{n=1}^N G(\mathbf{f}_{\text{fuse}})^n \right) + \text{CV}^2(\text{Load}(\mathbf{G})) \quad (3.12)$$

其中  $\text{CV}^2(\cdot)$  表示变异系数的平方，定义为样本方差除以样本均值的平方， $\sum_{n=1}^N G(f_{\text{fuse}}^n)$  统计每个专家在当前批次中所获得的门控权重之和（即重要性）， $\text{Load}(G)$  统计每个专家在当前批次中被选中的次数。当所有专家被等概率地选择且获得相同的累积权重时，两项变异系数均趋近于零，损失达到最小值。在实际实现中，编码器各层级（前三个层级）的 VPRM 模块各自计算负载均衡损失，并将其累加作为总的  $\mathcal{L}_{\text{Balance}}$ 。超参数  $\gamma$  作为平衡权重，控制负载均衡正则化的强度，在本章实验中设定为  $\gamma = 0.01$ ，以确保负载均衡约束不会过度干扰主重建任务的优化方向。

### 3.3 实验设置

为验证本章所提出的 VP-Net 在一体化 GISR 任务中的有效性，我们在三个具有代表性的 GISR 子任务上进行了全面的实验评估，分别为全色锐化 (Pansharpening)、深度图超分辨率 (Depth Image SR) 以及磁共振图像超分辨率 (MR Image SR)。本节将依次介绍实验所用的数据集、评测指标以及训练实现细节。

#### 3.3.1 数据集

本章实验涵盖了三类不同模态的 GISR 子任务，每类任务均选取了领域内广泛使用的公开基准数据集。图 3.5 展示了三种任务的输入与输出示例，包括低分辨率输入、高分辨率引导图像及高分辨率真值。

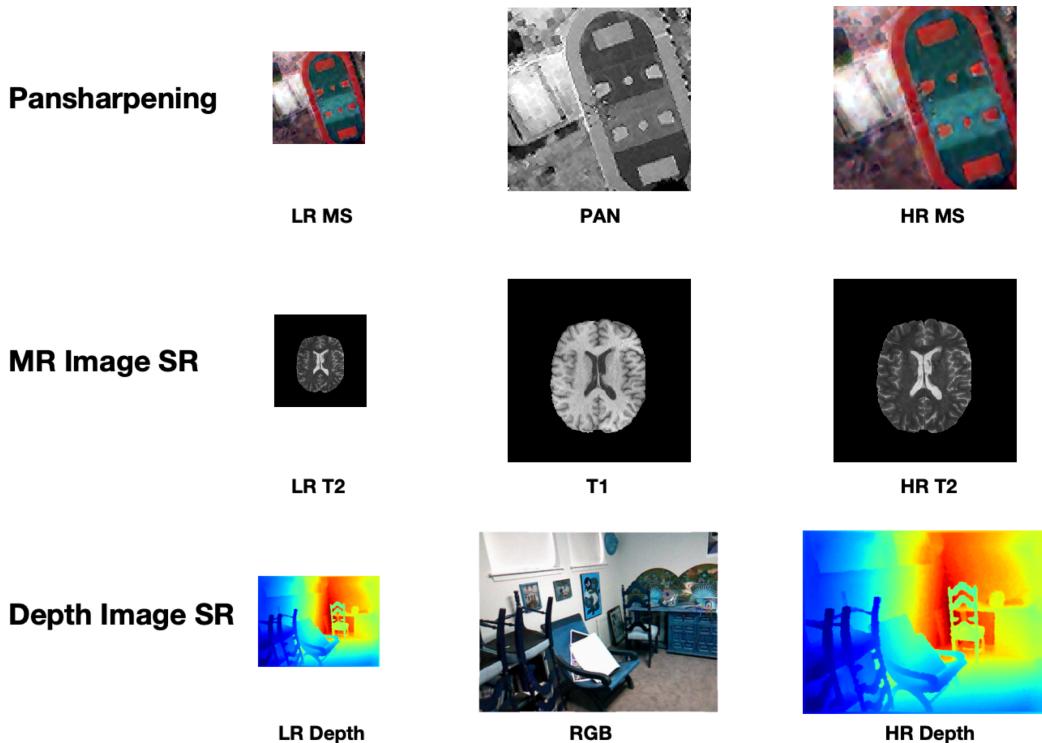


图 3.5 三种 GISR 子任务的数据集示例。每一行展示一个任务（从上至下依次为：全色锐化、磁共振图像超分、深度图超分），包括低分辨率输入、高分辨率引导图像及高分辨率真值。

对于全色锐化任务，我们在三个不同卫星传感器的数据集上进行实验，分别为 QuickBird (QB)、WorldView-4 (WV4) 和 GaoFen-1 (GF1)。这三个数据集在光谱响应特性、空间分辨率以及成像条件上各具特点，能够全面评估方法在不同遥感场景下的泛化能力。在数据预处理阶段，我们将原始图像对裁剪为小尺寸的训练样本，每组训练样本包含一幅  $128 \times 128 \times 1$  的高分辨率全色 (PAN) 图像作为引导图像、一幅  $32 \times 32 \times 4$  的低分辨率多光谱 (LRMS) 图像作为待超分的目标图像，以及一幅  $128 \times 128 \times 4$  的高分辨率多光谱 (HRMS) 图像作为训练的真实标签 (Ground Truth)。超分辨率倍率为  $4 \times$ 。

对于深度图超分辨率任务，我们选取了广泛认可的 NYU v2 基准数据集。该数据集包含 1449 对 RGB-D 图像，其中 RGB 图像作为高分辨率引导图像，深度图作为待超分的低分辨率目标图像。按照领域内通用的实验设置，我们使用其中 1000 对图像用于训练，剩余 449 对图像用于测试。为了全面评估方法在不同降质程度下的重建能力，我们分别设置了  $4 \times$ 、 $8 \times$  和  $16 \times$  三种超分辨率倍率。低分辨率深度图通过对原始深度图进行双三次下采样获得。

对于磁共振图像超分辨率任务，我们在 BrainTS 数据集上进行评估。该数据集包含 285 个多对比度磁共振体数据（Multi-contrast MR Volumes），涵盖了 T1、T1ce、T2 和 FLAIR 四种对比度模态。在数据准备过程中，我们从每个体数据中选取中间 100 个切片，并剔除内容信息不足的切片，最终获得 20480 个切片用于训练，2320 个切片用于测试。引导图像为同一受试者的另一对比度磁共振切片，超分辨率倍率分别设置为  $2 \times$ 、 $4 \times$  和  $8 \times$ 。

### 3.3.2 评测指标

针对不同任务的特点，我们采用了与之相适配的图像质量评价（Image Quality Assessment, IQA）指标。

在全色锐化任务中，我们采用三种互补的评价指标：峰值信噪比（Peak Signal-to-Noise Ratio, PSNR）用于衡量重建图像与真实标签之间的整体像素级保真度，其值越高表示重建质量越好；光谱角度映射（Spectral Angle Mapper, SAM）从光谱一致性的角度评估重建图像是否保持了原始多光谱图像的光谱特征，其值越低表示光谱失真越小；综合无量纲全局误差（Erreur Relative Globale Adimensionnelle de Synthèse, ERGAS）则从全局角度综合评价空间与光谱质量，其值越低表示整体重建质量越高。

在深度图超分辨率任务中，我们采用均方根误差（Root Mean Square Error, RMSE）作为主要评价指标。RMSE 直接度量重建深度图与真实深度图之间的像素级偏差，能够反映深度值估计的精确程度，其值越低表示重建精度越高。

在磁共振图像超分辨率任务中，我们采用 PSNR 和结构相似性指数（Structural Similarity Index Measure, SSIM）两种指标对重建质量进行评估。PSNR 衡量像素级的重建精度，SSIM 则从亮度、对比度和结构三个维度综合评价重建图像与真实标签之间的感知相似性。两项指标的值越高，均表示重建质量越好。

### 3.3.3 实验细节

所有实验均基于 PyTorch 深度学习框架实现，在单块 NVIDIA RTX 3090 GPU 上进行训练和测试。模型训练采用 Adam 优化器，批次大小（Batch Size）设置为

4, 初始学习率设定为  $4 \times 10^{-4}$ , 并采用余弦退火 (Cosine Annealing) 学习率调度策略在训练过程中逐步衰减学习率。

在骨干网络的架构配置方面, 编码器-解码器的四个层级分别包含  $L_1 = 3$ 、 $L_2 = 4$ 、 $L_3 = 4$ 、 $L_4 = 5$  个 Transformer Block, 体现了由浅到深逐步增加模型容量的设计理念。基础通道维度  $C$  设定为 42, 各层级的注意力头数分别为 1、2、4、8, 门控前馈网络的通道膨胀因子设为 2.66。在 VPRM 模块中, 专家总数  $M = 4$ , 每个像素激活的专家数  $k = 2$ 。负载均衡损失的权重系数  $\gamma = 0.01$ 。

在一体化训练设置中, 模型在三个任务的所有数据集上进行联合训练。由于不同任务的数据集规模存在差异, 我们将每个训练轮次 (Epoch) 的长度对齐至最小数据集 (深度图超分辨率任务的 1000 对训练样本)。在每个轮次结束后, 所有数据集均进行重新随机打乱 (Reshuffle), 以确保训练过程中的样本多样性。模型总共训练 500 个轮次。在单任务训练设置 (One-by-One) 中, 模型分别在各任务的数据集上独立训练, 每个任务同样训练 500 个轮次, 以便与一体化设置进行公平对比。

## 3.4 实验结果

为了验证本章提出的面向一体化任务的视觉感知与动态路由 GISR 方法 (VP-Net) 的有效性, 我们在全色锐化、磁共振图像超分及深度图超分三个典型任务上进行了广泛的实验。实验主要包含三个部分: 首先通过对比实验, 验证 VP-Net 在多任务混合训练设置下相较于基准模型及通用模型的性能优势; 其次通过消融实验, 深入探究视觉感知路由模块 (VARM) 的核心贡献及架构设计选择; 最后通过可视化分析, 直观展示动态路由机制在特征空间中的工作机理。

### 3.4.1 对比实验

#### All-in-One 方法对比

为了验证本章方法在多任务联合训练模式下的有效性, 我们将 VP-Net 与近年来提出的主流一体化图像复原网络进行了全面对比, 对比方法包括 Gridformer、Transweather、CAPTNet、AdaIR 以及 PromptIR。这些方法同样旨在

通过单一网络处理多种复原任务，是检验 VP-Net 以及动态路由机制能否缓解任务干扰的理想基准。

表 1 展示了全色锐化任务在 WV4、QB 和 GF1 三个卫星数据集上的定量评估结果。可以看出，VP-Net 在绝大多数指标上均取得了最优 (Bold) 的成绩。特别是与同样基于 Transformer 架构的 PromptIR 相比，VP-Net 在保持参数量相当的情况下，PSNR 指标得到进一步提升，这得益于视觉感知路由对不同光谱特征的精细化处理。

表 2 和 表 3 分别列出了磁共振图像超分辨率和深度图超分辨率任务的对比数据。在 MR 图像超分中，VP-Net 在  $2\times$ 、 $4\times$  和  $8\times$  三个尺度上均表现出一致的性能优势。在深度图超分任务中，VP-Net 的 RMSE 误差始终低于或持平于最强基准 PromptIR，尤其是在高倍率 ( $16\times$ ) 重建中，能够更好地重构深度结构。实验结果表明，通过动态路由实现特征的物理解耦，能够有效避免不同模态任务之间的负迁移现象，从而打破整体系统的性能瓶颈。

表 3.1 不同 all-in-one 方法在 Pansharpening 任务上的对比结果，最优结果用粗体表示，次优结果用下划线表示

Method	WV4			QB			GF1		
	PSNR↑	SAM↓	ERGAS↓	PSNR↑	SAM↓	ERGAS↓	PSNR↑	SAM↓	ERGAS↓
Gridformer	40.49	1.42	1.19	47.72	0.96	0.70	45.59	0.98	0.97
Transweather	39.54	1.67	1.39	46.26	1.14	0.80	43.99	1.22	1.18
CAPTNet	41.10	1.38	1.15	47.92	0.94	0.68	48.94	0.70	0.77
AdaIR	<u>43.46</u>	<u>1.06</u>	<u>0.85</u>	<u>49.90</u>	<b>0.75</b>	<b>0.54</b>	51.80	<u>0.47</u>	<u>0.54</u>
PromptIR	43.34	1.08	0.88	49.76	<u>0.76</u>	<u>0.55</u>	52.20	<b>0.46</b>	<b>0.53</b>
VPNet	<b>43.71</b>	<b>1.03</b>	<b>0.83</b>	<b>50.00</b>	<b>0.75</b>	<b>0.54</b>	<b>52.24</b>	<b>0.46</b>	<u>0.54</u>

表 3.2 不同 all-in-one 方法在 MR Image SR 任务上的对比结果，最优结果用粗体表示，次优结果用下划线表示

Method	2x		4x		8x	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Gridformer	42.15	0.9890	36.70	0.9662	33.64	0.9430
Transweather	39.91	0.9821	36.25	0.9594	33.73	0.9332
CAPTNet	40.43	0.9872	35.05	0.9573	32.09	0.9249
AdaIR	<u>44.27</u>	<b>0.9928</b>	38.92	0.9780	35.56	<u>0.9585</u>
PromptIR	44.16	<u>0.9926</u>	<u>38.97</u>	<u>0.9781</u>	<u>35.58</u>	0.9582
VPNet	<b>44.34</b>	<b>0.9928</b>	<b>39.02</b>	<b>0.9783</b>	<b>35.59</b>	<b>0.9589</b>

表 3.3 不同 all-in-one 方法在 Depth Image SR 任务上的对比结果，最优结果用粗体表示，次优结果用下划线表示

Method	X4 RMSE↓	X8 RMSE↓	X16 RMSE↓
Gridformer	1.67	3.04	5.72
Transweather	2.97	4.14	6.20
CAPTNet	1.64	2.88	5.13
AdaIR	1.53	<u>2.79</u>	5.09
PromptIR	<u>1.48</u>	<b>2.69</b>	<u>4.89</u>
VPNet	<b>1.47</b>	<b>2.69</b>	<b>4.88</b>

### One-by-One 方法对比

除了验证一体化训练的性能外，我们还在单任务独立训练（One-by-One）设置下评估了 VP-Net 的潜力，旨在探究其作为通用骨干网络是否具备与针对特定任务设计的专家模型相抗衡的能力。

在全色锐化任务中（表 4），我们将 VP-Net 与当前领域内的专用 SOTA 方法进行了对比，包括 AWFLN、DISPNet、LAGConv、M3DNet、FusionMamba 以及基于扩散模型的 DifPan。结果显示，VP-Net 的性能优于大多数专用模型（如 FusionMamba），仅略逊于计算代价极高的扩散模型 DifPan。这表明即便不依赖特定任务的先验设计，仅凭强大的骨干网络与动态路由机制，VP-Net 也能达到领域前沿水平。

在磁共振图像超分任务中（表 5），我们引入了 MASA、SANet 和 DuDoNet 等专用模型作为对比。结果表明，VP-Net 在  $2\times$  和  $8\times$  倍率下的重建质量（PSNR 分别为 44.98 dB 和 35.47 dB）甚至超过了表现优异的 DuDoNet，在  $4\times$  倍率下也与之相当，体现了模型在从医学影像中提取精细解剖结构方面的强大能力。

在深度图超分任务中（表 6），VP-Net 与 GeoDSR、DKN、SGNet 等方法进行了比较。在高倍率( $16\times$ )条件下，VP-Net 取得了最低的 RMSE(4.67)，优于最新的 SGNet(4.77)，证明了模型在处理极低分辨率输入时的鲁棒性。这些结果证实了 VPRM 模块不仅在多任务冲突场景下有效，在单任务场景下也能通过感知输入图像的视觉特性自适应地增强特征表达，具有良好的泛化性。

表 3.4 不同 one-by-one 方法在 Pansharpening 任务上的对比结果, 最优结果用粗体表示, 次优结果用下划线表示

Method	WV4			QB			GF1		
	PSNR↑	SAM↓	ERGAS↓	PSNR↑	SAM↓	ERGAS↓	PSNR↑	SAM↓	ERGAS↓
AWFLN	42.13	1.23	0.98	49.14	0.82	0.60	49.70	0.62	0.64
DISPNet	40.79	1.43	1.15	48.18	0.93	0.70	46.61	0.86	0.84
LAGConv	41.54	1.33	1.07	47.89	0.96	0.71	47.66	0.79	0.79
M3DNet	42.27	1.20	0.97	49.45	0.80	0.58	49.80	0.62	0.64
FusionMamba	42.84	<u>1.15</u>	<u>0.92</u>	49.50	0.81	0.59	50.79	<u>0.55</u>	0.60
DifPan	<b>43.89</b>	<b>1.03</b>	<b>0.83</b>	<b>50.22</b>	<b>0.72</b>	<b>0.53</b>	<b>52.33</b>	<b>0.46</b>	<b>0.51</b>
VPNet	<u>43.57</u>	<b>1.06</b>	<b>0.87</b>	<u>49.77</u>	<u>0.75</u>	<u>0.56</u>	<u>52.14</u>	<b>0.50</b>	<u>0.56</u>

表 3.5 不同 one-by-one 方法在 MR Image SR 任务上的对比结果, 最优结果用粗体表示, 次优结果用下划线表示

Method	2x		4x		8x	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
MGDUN	40.97	<u>0.9893</u>	35.34	0.9632	32.61	0.9361
MADUNet	38.94	0.9843	34.09	0.9548	30.18	0.9114
MINet	38.84	0.9788	35.62	0.9625	32.38	0.9335
MASA	40.37	0.9883	34.93	0.9590	30.41	0.8954
McMRSR	40.96	<u>0.9893</u>	35.20	0.9596	30.55	0.8985
SANet	42.36	0.9881	36.56	0.9583	33.53	0.9419
DuDoNet	<u>44.96</u>	<b>0.9919</b>	<b>38.99</b>	<b>0.9744</b>	<u>35.43</u>	<u>0.9533</u>
VPNet	<b>44.98</b>	<b>0.9919</b>	<u>38.95</u>	<u>0.9742</u>	<b>35.47</b>	<b>0.9550</b>

表 3.6 不同 one-by-one 方法在 Depth Image SR 任务上的对比结果, 最优结果用粗体表示, 次优结果用下划线表示

Method	4x		8x		16x	
	RMSE↓	RMSE↓	RMSE↓	RMSE↓	RMSE↓	RMSE↓
GeoDSR	1.42		<u>2.62</u>		4.86	
DAGF	1.36		2.87		6.06	
AHMF	1.40		2.89		5.64	
DCTNet	1.59		3.16		5.84	
DKN	1.62		3.26		6.51	
FDKN	1.86		3.58		6.69	
SGNet	<b>1.10</b>		<b>2.44</b>		<u>4.77</u>	
VPNet	<u>1.32</u>		<b>2.44</b>		<b>4.67</b>	

### 3.4.2 消融实验

为了验证 VP-Net 中核心设计选择的合理性，我们进行了一系列消融实验。实验主要关注两个方面：一是视觉感知路由模块（VPRM）本身的有效性，二是 VPRM 在网络架构中的最佳部署位置。

#### VPRM 有效性消融分析

为了验证 VPRM 在缓解多任务干扰方面的贡献，我们构建了一个基准模型（w/o VPRM），即移除 VP-Net 中所有的 VPRM 模块，退化为标准的共享参数 Restormer 架构，并在相同的数据设置下进行多任务联合训练。表 7 展示了基准模型与完整 VP-Net 的性能对比。

由表可知，相较于基准模型，引入 VPRM 后的 VP-Net 在所有任务的各项指标上均取得了显著提升。例如，在全色锐化任务的 GF1 数据集上，PSNR 提升了近 1 dB；在深度图超分任务中，RMSE 误差也得到明显降低。这一结果充分表明，简单的参数共享策略在处理差异巨大的异构模态时存在局限性，而 VPRM 通过基于视觉特征的动态路由机制，成功实现了任务特征的物理解耦，有效缓解了负迁移现象，从而大幅提升了一体化模型的重建质量。

表 3.7 VPRM 模块有效性消融实验结果分析

Method	Pansharpening			MR Image SR			Depth Image SR		
	WV4	QB	GF1	2×	4×	8×	4×	8×	16×
w/o VPRM	43.23	49.55	51.29	43.77	38.03	34.87	1.58	2.85	4.97
VPNet	43.71	50.00	52.24	44.34	39.02	35.59	1.47	2.69	4.88

#### VPRM 位置消融分析

VPRM 应当部署在网络的什么位置才能最大化其效能？为了回答这个问题，我们对比了三种不同的部署策略：(1) 仅在编码器中部署（Encoder，即 VP-Net 最终方案）；(2) 仅在解码器中部署（Decoder）；(3) 同时在编码器和解码器中部署（Both）。实验结果如表 8 所示。

首先，对比 Encoder 和 Decoder 的结果可以发现，将路由模块置于编码阶段的性能明显优于解码阶段。这印证了我们的假设：不同任务间的模态差异（Domain

Gap) 主要存在于特征提取与理解阶段。在编码器中尽早进行特征分流与解耦, 能够避免特征在深层发生混淆, 为后续重建打下良好基础。

其次, 对比“Encoder”与“Both”方案, 可以观察到“Both”设置虽然在全色锐化和磁共振超分任务上取得了极其微弱的性能优势 (例如 Pan WV4 上 PSNR 仅高出 0.02 dB), 但在深度图超分任务上, 其性能甚至略逊于仅编码器设置。更关键的是, 双端部署带来了巨大的计算代价。引入更多的专家网络不仅显著增加了模型的参数量, 还大幅提高了推理延迟。此外, 过多的动态路由决策节点显著增加了训练的复杂性, 导致模型优化难度增大, 极难收敛。综合考虑性能、效率与训练稳定性, 我们认为仅在编码器中部署 VPRM 是最优的选择, 它在保持高效推理的同时, 实现了与更复杂模型相当的性能表现。

表 3.8 不同模块位置 (Encoder/Decoder) 对各任务性能的影响分析

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

### 3.4.3 动态路由机制的可视化分析

为了直观地揭示 VP-Net 如何缓解多任务干扰, 我们可视化了不同任务在编码器各层级 (L1-L4) 对各个专家 (E0-E3) 的平均激活率。激活热力图如图 3.6 所示, 其中颜色的深浅代表每个专家被激活的频率。

从图中可以观察到明显的专家选择偏好差异, 这证实了 VP-Net 成功实现了基于任务特性的特征解耦:

1. 浅层 (L1) 的显著分化: 在网络的浅层, 专家选择的差异最为剧烈。例如, MRI 任务极度依赖专家 E1 (深蓝色高亮), 而对其他专家 (特别是 E2 和 E3) 的激活率极低。相反, Pan 任务则表现出对专家 E2 和 E3 的偏好, 对 E0 和 E1 的依赖相对较少。这表明网络在特征提取的早期阶段就已经开始区分不同模态的视觉特征, 将光谱信息丰富的全色影像与解剖结构复杂的磁共振影像路由至完全不同的处理路径, 从而从源头上避免了特征混淆。

2. 深层 (L4) 的趋同性：随着网络层级的加深，不同任务的专家激活分布逐渐趋于均匀（颜色差异变浅）。这说明在深层语义空间中，尽管输入模态不同，网络提取的高层抽象特征开始表现出一定的共性（如对物体轮廓或纹理复原的通用需求）。此时，各任务开始共享更多的专家参数，实现了知识的迁移与互补。
3. 任务间的独特性与重叠：Depth 任务与 Pan 任务在某些专家（如 L1\_E1）上的激活模式存在重叠，但也保留了自身的独特性。这种“部分共享、部分独立”的路由机制正是 MoE 架构的优势所在——它既不像硬性参数隔离 (Hard Parameter Sharing) 那样完全阻断任务间的联系，也不像全参数共享那样导致剧烈的干扰，而是通过动态门控实现了一种“软性”的平衡，确保每个任务都能找到最优的参数组合。

综上所述，可视化分析证实了 VPRM 能够根据输入模态的底层视觉特征，自动学习并规划出差异化的特征处理路径。这种动态分流策略不仅在理论上解释了 VP-Net 性能提升的原因，也在实践中展示了其处理多模态冲突的强大能力。

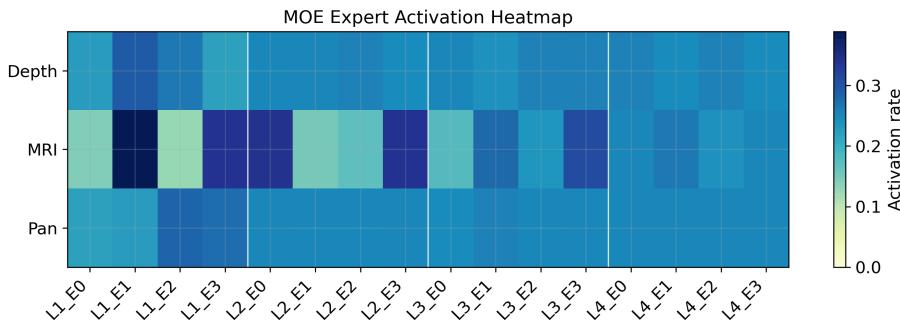


图 3.6 动态路由机制激活热力图

### 3.5 本章小结

本章针对多模态引导图像超分辨率任务中存在的“负迁移”与参数干扰问题，提出了一种基于视觉感知与动态路由的一体化网络模型 (VP-Net)。该方法通过引入视觉感知路由模块 (VPRM)，利用引导图像的纹理与结构信息作为“视觉罗盘”，实现了特征处理路径的动态规划与物理隔离。

首先，本章详细阐述了 VP-Net 的整体架构设计，重点介绍了 VPRM 模块如何通过稀疏门控机制在像素级别分配专家网络，从而解决不同模态数据在特征空间中的分布冲突。其次，通过在全色锐化、磁共振图像超分和深度图超分三个典型任务上的广泛实验，验证了 VP-Net 在一体化训练设置下相较于现有 SOTA 方法的性能优势。特别是在处理异构模态数据时，VP-Net 展现出了优异的鲁棒性与泛化能力。最后，通过消融实验与可视化分析，深入揭示了动态路由机制在特征解耦与知识共享中的工作机理，证明了“仅编码器部署”策略在性能与效率之间的最佳平衡。本章的研究为构建通用、高效的多模态图像复原系统提供了新的视角与解决方案。

## 第四章 融合多模态语义提示的一体化 GISR 方法

### 4.1 引言

在前一章中,为了解决一体化引导图像超分辨率(GISR)任务中存在的“参数干扰”与“负迁移”问题,我们尝试了一种基于纯视觉感知的解决方案——VP-Net。该方法创新性地引入了动态路由机制,利用引导图像的纹理、边缘等底层视觉特征作为隐式信号,将不同模态的数据分流至不同的专家网络进行处理。实验结果证实,这种基于视觉特征的“物理隔离”策略在一定程度上缓解了多任务间的优化冲突,优于传统的静态参数共享模型。

然而,VP-Net本质上是一个仅依赖数据驱动的“视觉主导”版本,可视作本章所提完整方法的一个退化特例。它虽然能够“看见”图像纹理的差异,却无法真正“理解”任务的语义本质。如图 图 4.1 所示,在处理复杂多变的 GISR 任务时,这种仅依赖底层视觉特征的路由机制暴露出了两个关键局限:

首先是视觉歧义性(Visual Ambiguity)。不同任务的图像在局部可能表现出极为相似的纹理统计特性(Domain Overlap)。例如,平滑的深度图区域与磁共振影像的背景区域在梯度分布上可能难以区分,单纯依靠视觉感知的路由模块极易产生混淆,导致无法精确激活最优的专家组合。

其次是语义缺失(Semantic Absence)。GISR 任务通常包含明确的先验定义,如数据源的传感器类型(QuickBird vs WorldView)、具体的任务目标(全色锐化 vs 深度恢复)以及缩放倍率( $4 \times$  vs  $8 \times$ )。这些高层语义信息对于指导模型进行针对性的重建至关重要(例如,全色锐化需要保持光谱一致性,而深度恢复关注几何结构),但在 VP-Net 中,这些关键的上下文信息被完全忽略了。

为了弥补这一从“感知”到“认知”的鸿沟,并将一体化模型的性能推向新的高度,本章在 VP-Net 的架构基础上,提出了一种融合多模态语义提示的一体化 GISR 方法(MAG-Net, Multi-modal All-in-one Guided Network)。如果说 VP-Net 是依靠直觉进行判断的“观察者”,那么 MAG-Net 则是不仅具备敏锐视觉,还能听懂明确指令的“执行者”。

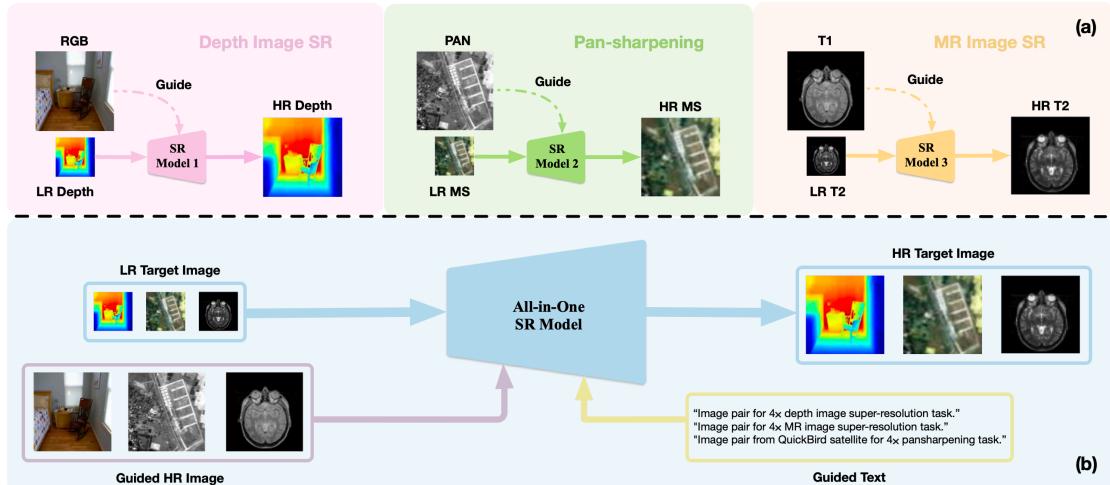


图 4.1 纯视觉感知面临的歧义性挑战与多模态语义提示的引入

MAG-Net 的核心思想在于引入文本语义先验 (Textual Semantic Prior) 来修正并增强动态路由的决策过程。为此，我们设计了一个多模态提示生成模块 (Multi-modal Prompt Generation Module, MPGM)。该模块利用预训练的大规模视觉-语言模型 (如 CLIP) 作为文本编码器，将任务的具体描述 (如“Pansharpening for satellite imagery”、“Scale factor 4x”) 转化为富含语义的高维特征向量，并将其作为显式的“语义锚点 (Semantic Anchors)”注入到动态路由网络中。

通过这种“视觉感知 + 语义引导”的双重驱动机制，MAG-Net 能够在像素级和任务级同时对特征进行精细化调控。高层语义提示为路由模块提供了全局的任务上下文，消除了视觉特征的歧义性，确保任务被正确分类；而底层视觉感知则保留了对局部空间纹理的自适应能力。两者相辅相成，使得模型不仅实现了任务间的彻底解耦，更能够利用不同任务间的语义关联促进知识的正向迁移。本章将详细阐述 MAG-Net 的网络架构与提示学习机制，并通过对比实验证明，在引入语义交互后，模型在全色锐化、深度图超分及磁共振超分三个子任务上均取得了显著优于 VP-Net 及现有专用模型的性能表现。

## 4.2 方法

本章这一部分将详细阐述 MAG-Net 的实施细节。针对多模态一体化任务中纯视觉感知存在的“歧义性”与“语义缺失”问题，MAG-Net 在 VP-Net 的架构基础上进行了语义增强。其核心在于引入了任务描述文本作为显式的先验信息，并通

过设计专门的模块将语义特征注入到动态路由的决策过程中,从而实现“视觉+语义”双重驱动的任务解耦。

#### 4.2.1 整体网络架构

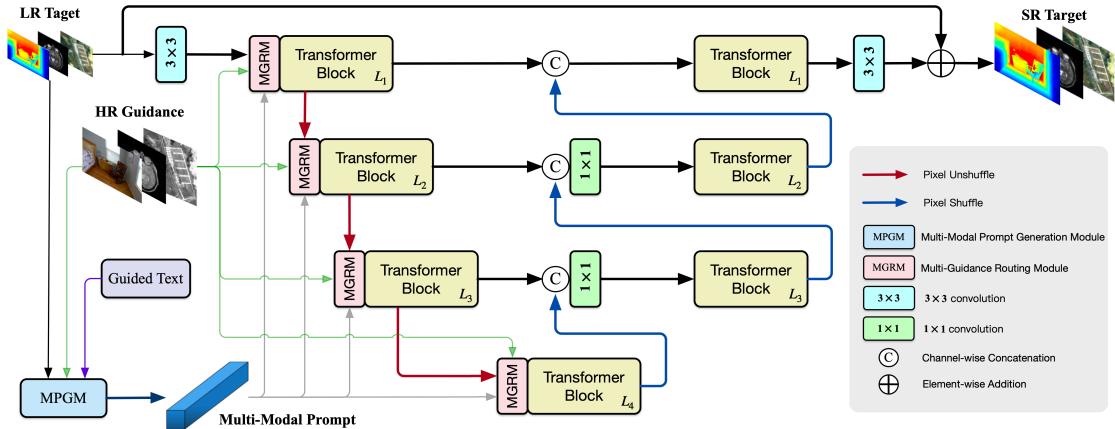


图 4.2 MAG-Net 网络整体架构示意图。模型接收低分辨率图像、高分辨率引导图像及任务语义描述作为输入,通过多模态提示生成模块 (MPGM) 提取语义先验,并结合视觉特征在 MGRM 中动态激活特定的专家网络

如图图 4.2 所示, MAG-Net 建立在一个分层的 U 型编码器-解码器骨干网络之上,整体流程包含三个关键阶段: 多模态输入编码、基于语义增强的动态特征映射、以及高分辨率图像重建。

多模态输入与浅层特征提取模型的输入由三部分组成: 待恢复的低分辨率目标图像  $I_{LR} \in \mathbb{R}^{H \times W \times C_{in}}$ 、同场景的高分辨率引导图像  $I_{HR} \in \mathbb{R}^{sH \times sW \times C_{guide}}$  ( $s$  为超分倍率), 以及描述当前任务属性的自然语言文本  $T$  (例如 “Pansharpening for satellite imagery” 或 “Depth map super-resolution x8”)。针对图像数据, 我们利用两个独立的  $3 \times 3$  卷积层分别将  $I_{LR}$  和  $I_{HR}$  映射到统一的特征通道维度  $C$ , 得到浅层特征  $F_{LR}^0$  和  $F_{HR}^0$ 。其中  $F_{LR}^0$  作为主干网络的输入流,  $F_{HR}^0$  则作为视觉引导信号被送入后续的路由模块。针对文本数据  $T$ , 我们设计了多模态提示生成模块 (MPGM)。该模块利用预训练的视觉-语言大模型 (如 CLIP) 作为文本编码器, 将离散的文本指令转化为连续的高维语义提示向量  $P_{sem}$ 。这个语义向量承载了任务的全局定义 (如传感器类型、目标模态), 充当了后续动态路由过程中的“语义锚点”。

融合多模态引导的编码器编码器包含 4 个层级，旨在逐步提取深层抽象特征。每个层级堆叠了若干 Transformer Block (基于 MDTA 和 GDFN)，用于捕捉图像的局部细节与长距离依赖。与 VP-Net 仅依赖视觉特征进行路由不同，MAG-Net 在编码器的每个层级嵌入了多模态引导路由模块 (Multi-modal Guided Routing Module, MGRM)。在第  $l$  个层级，MGRM 同时接收三组输入信号：当前主干特征  $F_l^{\text{enc}}$ 、下采样的视觉引导特征  $F_{\text{guide}}^l$  以及全局语义提示  $P_{\text{sem}}$ 。MGRM 作为一个智能的“调度中心”，通过交叉注意力或特征拼接的方式，融合底层的视觉纹理信息与高层的语义任务指令，生成像素级的门控权重。这些权重动态地激活特定的专家网络 (Experts) 组合，从而对特征进行针对性的处理。通过引入  $P_{\text{sem}}$ ，MGRM 能够有效消除仅靠视觉特征难以区分的歧义（例如，区分纹理相似但物理属性不同的深度图平滑区与 MRI 背景区），确保特征被路由至语义上正确的专家路径，实现了比 VP-Net 更彻底的任务解耦。层级间的下采样操作采用 Pixel-Unshuffle，以在降低分辨率的同时保留完整的通道信息。

解码器与图像重建解码器结构与 VP-Net 保持一致，包含 4 个对称的层级。由于编码阶段的 MGRM 已经完成了基于语义的任务解耦与特征增强，解码阶段主要负责利用这些纯净的特征恢复空间细节。因此，解码器采用标准的共享参数 Transformer Block，未引入额外的路由机制。解码器通过跳跃连接 (Skip Connections) 融合编码器传递的多尺度特征，并通过 Pixel-Shuffle 操作逐步恢复图像的空间分辨率。最终，网络输出残差图像  $I_{\text{res}}$ ，并与经双线性插值上采样的输入  $I_{\text{LR}}$  相加，得到最终的超分辨率结果  $I_{\text{SR}}$ ：

$$I_{\text{SR}} = I_{\text{res}} + \text{Upsample}(I_{\text{LR}}) \quad (4.1)$$

综上所述，MAG-Net 通过在骨干网络中无缝集成文本语义先验，将“所见” (Visual Perception) 与“所知” (Semantic Cognition) 相结合，构建了一个具备认知能力的一体化图像修复框架。

#### 4.2.2 多模态提示生成模块

为了向网络提供精确且对任务敏感的语义引导，我们需要将离散的自然语言描述转化为能够与视觉特征进行交互的连续嵌入向量。然而，直接使用预训练语言模型（如 BERT 或 CLIP）输出的特征往往存在两个问题：一是领域偏差

(Domain Gap)，通用大模型对特定遥感或医学术语的理解可能不够细粒度；二是模态隔离，纯文本特征无法感知当前输入图像的具体状态（如噪声水平或纹理复杂度）。为此，我们设计了多模态提示生成模块（MPGM），引入了基于字典学习（Dictionary Learning）的思想。

MPGM 的处理流程如图 4.3 所示，包含视觉-文本特征提取、跨模态特征调制以及基于字典的提示重构三个阶段。

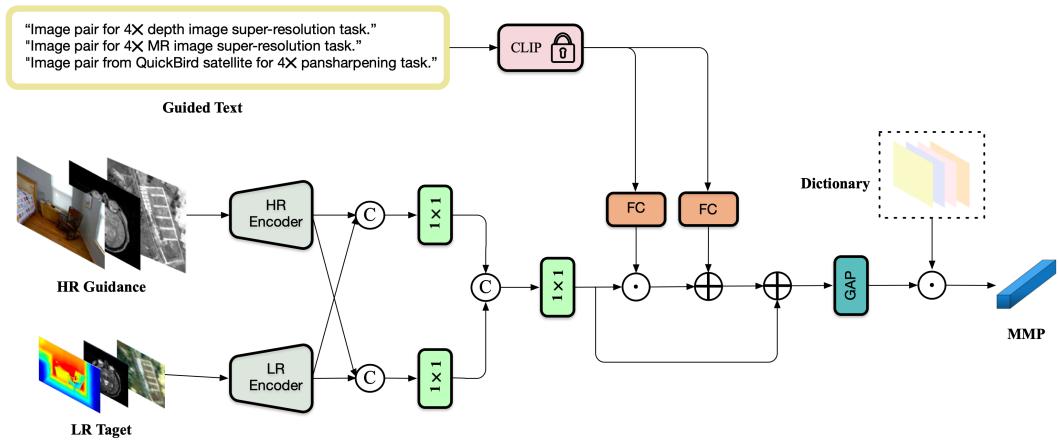


图 4.3 多模态提示生成模块（MPGM）处理流程示意图

1. 视觉-语言特征提取首先，对于描述任务属性的文本指令  $T$ （例如“PanSharpening with 4 spectral bands”），我们利用预训练的 CLIP 文本编码器提取其高维语义特征  $F_{\text{text}} \in \mathbb{R}^{D_t}$ 。与此同时，为了确保生成的提示能适应当前的图像内容（Context-Awareness），我们提取图像的视觉上下文。通过两个轻量级的卷积网络  $\mathcal{E}_{\text{LR}}$  和  $\mathcal{E}_{\text{HR}}$  分别提取低分辨率模型输入  $I_{\text{LR}}$  和高分辨率引导图  $I_{\text{HR}}$  的浅层特征，并通过通道拼接与  $1 \times 1$  卷积融合操作获得联合视觉特征  $F_{\text{vis}} \in \mathbb{R}^{C \times H \times W}$ ：

$$F_{\text{vis}} = \mathcal{F}_{\text{fuse}}(\text{Concat}(\mathcal{E}_{\text{LR}(I_{\text{LR}})}, \mathcal{E}_{\text{HR}(I_{\text{HR}})})) \quad (4.2)$$

这使得模块在处理不同图像时能够具有自适应的感知能力。

2. 跨模态特征调制为了将显式的语义指令注入到底层的视觉特征中，我们采用特征线性调制（FiLM）机制。具体而言，利用两个全连接层从文本特征  $F_{\text{text}}$  中预测出缩放系数  $\gamma$  和平移系数  $\beta$ ，对视觉特征  $F_{\text{vis}}$  进行通道级的仿射变换：

$$F_{\text{mix}} = (1 + \gamma(F_{\text{text}})) \odot F_{\text{vis}} + \beta(F_{\text{text}}) \quad (4.3)$$

其中  $\odot$  表示逐元素乘法，公式中的 1 代表残差连接，保证了视觉信息的完整传递。经过这一步，生成的  $F_{\text{mix}}$  既包含了图像的空间纹理信息，也被赋予了明确的任务语义倾向。

3. 基于字典的提示重构为了获得更加紧凑且适用于 GISR 任务的提示表示，我们不再直接使用混合特征，而是将其作为一个“查询信号”，去检索一组可学习的语义锚点。我们预定义了一个任务无关的共享提示字典  $\mathcal{D} = \{\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_K\}$ ，其中  $\mathbf{d}_k \in \mathbb{R}^D$  为第  $k$  个潜在的语义原子 (Semantic Atom)， $K$  为原子总数。我们将  $F_{\text{mix}}$  进行全局平均池化 (GAP) 得到向量  $z$ ，并通过一个线性分类器预测其在字典上的注意力分布（即当前样本属于哪种潜在任务模式的概率）：

$$\mathbf{w} = \text{Softmax}(\mathbf{W}_p z) \in \mathbb{R}^K \quad (4.4)$$

最终的语义提示  $P_{\text{sem}}$  由字典原子加权组合而成：

$$P_{\text{sem}} = \sum_{k=1}^K w_k \mathbf{d}_k \quad (4.5)$$

这种设计即是一种“软聚类”过程，模型自动学习从复杂的视觉-文本混合空间到一组纯净的任务基向量的映射。由此生成的  $P_{\text{sem}}$  既具有文本赋予的语义指向性，又经过了图像内容的校准，为后续的动态路由提供了鲁棒的先验信号。

### 4.2.3 融合多模态提示的动态路由模块

在第三章中，我们提出了视觉感知路由模块 (VPRM)，通过感知引导图像的纹理和边缘信息来实现实任务特征的物理解耦。然而，正如引言所述，仅依靠视觉特征容易产生歧义。为了解决这一问题，我们将多模态提示生成模块 (MPGM) 生成的语义提示  $P_{\text{sem}}$  引入到路由决策中，设计了融合多模态提示的动态路由模块 (Multi-modal Guided Routing Module, MGRM)。

如图图 10 所示，MGRM 在架构上继承了 VPRM 的稀疏门控混合专家 (MoE) 设计，保留了“混合专家组” (Mixture of Experts) 和“稀疏分发与聚合”机制 (详见第 3.2.2 节)。其核心改进在于门控网络 (Gating Network) 的输入特征构造，从单纯的“视觉感知”升级为“视觉-语义联合感知”。

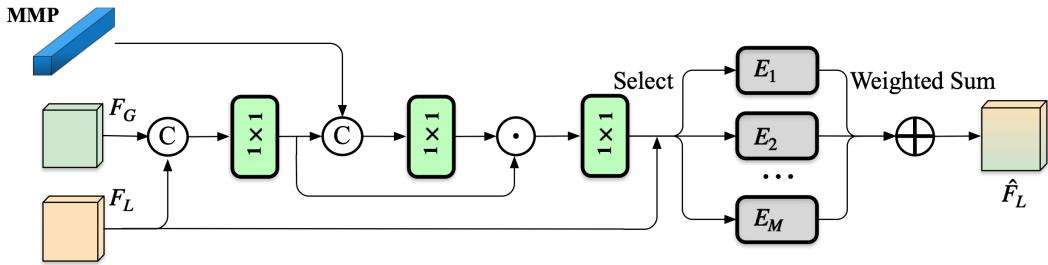


图 4.4 融合多模态提示的动态路由模块 (MGRM) 结构示意图

语义增强的门控机制 (Semantically Enhanced Gating) 在编码器的第  $l$  层级，MGRM 接收三路输入：主干特征  $F_l$ 、引导特征  $F_{\text{guide}}$  以及语义提示向量  $P_{\text{sem}}$ 。与 VPRM 仅拼接  $F_l$  和  $F_{\text{guide}}$  不同，MGRM 采用了一种级联融合策略来生成门控信号：

首先，主干特征与引导特征在通道维度拼接，并经过一个  $1 \times 1$  卷积层进行初步融合，得到视觉上下文特征  $F_{\text{vis}'}$ ：

$$F_{\text{vis}'} = \text{Conv}_1([\text{Concat}(F_l, F_{\text{guide}})]) \quad (4.6)$$

随后，为了将全局语义信息注入到局部门控决策中，我们将语义提示向量  $P_{\text{sem}}$  在空间维度上进行广播 (Broadcast)，使其尺寸扩充为与  $F_{\text{vis}'}$  一致，即  $P_{\text{sem}} \in \mathbb{R}^{B \times C \times H \times W}$ 。扩充后的语义提示与视觉上下文特征再次拼接，并通过第二层卷积进行深度融合：

$$F_{\text{gate}} = \text{Conv}_2([\text{Concat}(F_{\text{vis}'}, P_{\text{sem}})]) \quad (4.7)$$

这一步至关重要。通过引入  $P_{\text{sem}}$ ，门控网络不仅能感知像素局部的纹理差异（由  $F_{\text{guide}}$  提供），还能明确知晓当前任务的全局定义（由  $P_{\text{sem}}$  提供）。例如，在处理深度图超分任务时，即便某些区域的纹理与 MRI 图像相似，由于  $P_{\text{sem}}$  中包含了明确的“Depth Estimation”语义编码，门控网络依然能够抑制 MRI 相关专家的激活，强制将特征路由至深度恢复专家。

最后，融合后的特征  $F_{\text{gate}}$  经过激活函数调制，并与其自身产生的注意力图相乘（类似 GLU 门控线性单元结构），生成最终用于预测路由分数的像素级特征向量  $f_{\text{gate}}$ ：

$$f_{\text{gate}} = \text{Conv}_3(\text{GELU}(F_{\text{gate}}) \odot F_{\text{vis}'}) \quad (4.8)$$

基于该特征，我们沿用第三章所述的带噪声 Top-k 门控机制（Noisy Top-k Gating），计算每个像素分配给各个专家的权重，并进行稀疏分发。

$$G(\mathbf{f}_{\text{gate}}) = \text{Softmax}(\text{TopK}(\mathbf{f}_{\text{gate}} W_g + \text{Noise})) \quad (4.9)$$

通过这种设计，MGRM 实际上构建了一个条件概率模型  $P(\text{Expert} | \text{Visual}, \text{Semantic})$ 。相较于 VP-Net 的  $P(\text{Expert} | \text{Visual})$ ，MGRM 在决策时引入了额外的语义条件变量，大大降低了路由选择的不确定性 (Uncertainty)，实现了更精准、更鲁棒的一体化任务解耦。

### 4.3 实验设置

### 4.4 实验结果

#### 4.4.1

### 4.5 本章小结

## 第五章 一体化引导图像超分辨率系统

5.1 引言

5.2 需求分析

5.3 系统设计

5.4 开发环境和依赖

5.5 本章小结

## 第六章 总结与展望

6.1 工作总结

6.2 未来展望

## 参考文献

- [1] 蒋有绪, 郭泉水, 马娟, 等. 中国森林群落分类及其群落学特征[M]. 北京: 科学出版社, 1998: 11-12.
- [2] 中国力学学会. 第3届全国实验流体力学学术会议论文集[C]. 天津: \*\*出版社, 1990: 20-24.
- [3] World Health Organization. Factors Regulating the Immune Response: Report of WHO Scientific Group[R]. Geneva, 1970.
- [4] 张志祥. 间断动力系统的随机扰动及其在守恒律方程中的应用[D]. 北京, 1998: 50-55.
- [5] 河北绿洲生态环境科技有限公司. 一种荒漠化地区生态植被综合培育种植方法:中国, 01129210.5: 1129210.5[P/OL]. 2001. <http://211.152.9.47/sipoasp/zlijs/hyjs-yxnew.%20asp?recid=01129210.5&leixin>.
- [6] 国家标准局信息分类编码研究所. 世界各国和地区名称代码[Z]. 1986.
- [7] 李炳穆. 理想的图书馆员和信息专家的素质与形象[J]. 图书情报工作, 2000, 2000(2): 5-8.
- [8] 丁文祥. 数字革命与竞争国际化[J]. 中国青年报, 2000(15).
- [9] 江向东. 互联网环境下的信息处理与图书管理系统解决方案[J/OL]. 情报学报, 1999, 18(2): 4. <http://www.chinainfo.gov.cn/periodical/gbxb/gbxb99/gbxb990203>.
- [10] CHRISTINE M. Plant physiology: plant biology in the Genome Era[J/OL]. Science, 1998, 281: 331-332. <http://www.sciencemag.org/cgi/collection/anatmorp>.

## 致谢

感谢以下模板提供的参考：

- [modern-nju-thesis](#) by [OrangeX4](#)
- [ECNU-Undergraduate-LaTeX](#) by [Yijun Yuan](#)
- [华东师范大学硕士论文模板-2023](#) by [ivyee17](#)
- [ECNU\\_graduation\\_thesis\\_template](#) by [ECNU-ICA](#)
- [ECNU-Dissertations-Latex-Template](#) by [Karl Xing](#)

## 附录

### 7.1 附录子标题

#### 7.1.1 附录子子标题

附录内容，这里也可以加入图片，例如图 7.1。



图 7.1 图片测试

## 攻读硕/博士学位期间科研情况

- [1] J. von Neumann, “First draft of a report on the EDVAC,” IEEE Annals of the History of Computing, vol. 15, no. 4, pp. 27–75, 1993, doi: 10.1109/85.238389.
- [2] A. M. Turing, “On Computable Numbers, with an Application to the Entscheidungsproblem,” Proceedings of the London Mathematical Society, vol. s2-42, no. 1, pp. 230–265, 1937, doi: 10.1112/plms/s2-42.1.230.