arXiv:2512.18660v1 [cs.CV] 21 Dec 2025

PMPGuard：遥感图像-文本检索中的伪匹配对识别

Pengxiang Ouyang[1],**Qing Ma**[2*], **Zheng Wang**[1],**Cong Bai**[1,3],[1]浙江工业大学计算机科学与技术学院，中国杭州 310023[2]浙江工业大学数学科学学院，中国杭州 310023[3]浙江省视觉信息智能处理重点实验室，中国杭州 310023 oypx@zjut.edu.cn, maqing@zjut.edu.cn,zhengwang@zjut.edu.cn, congbai@zjut.edu.cn

摘要

由于存在伪匹配对（PMP），即语义不匹配或对齐较弱的图文对，遥感（RS）图像-文本检索在真实世界数据集上面临重大挑战，这些伪匹配对阻碍了可靠跨模态对齐的学习。为解决这一问题，我们提出了一种新颖的检索框架，利用跨模态门控注意力和正负感知注意力机制，以减轻此类噪声关联的影响。门控模块动态调节跨模态信息流，而感知机制则在对齐学习过程中明确区分有用的（正向）线索与误导性的（负向）线索。在三个基准遥感数据集（即 RSICD、RSITMD 和 RS5M）上进行的大量实验表明，我们的方法始终取得最先进的性能，彰显了其在处理真实世界中的不匹配以及遥感图像-文本检索任务中的伪匹配对方面的稳健性和有效性。

引言

遥感（RS）图像—文本检索旨在建立航空影像与自然语言描述之间准确的语义对应关系（Li、Ma 和 Zhang，2021年；Qu 等，2016年），支撑从灾害监测、土地利用分类到城市规划等广泛的地球观测应用。该任务本质上是跨模态的：给定文本查询，必须检索出最相关的RS图像，反之亦然。然而，若干挑战阻碍了相关研究的进展。首先，视觉域与文本域之间存在显著的模态鸿沟，而俯视场景通常具有丰富且细粒度的语义，这进一步加剧了这一问题（Ouyang 等，2024年；Lin 等，2025年；Ouyang、Ma 和 Bai，2025年）。其次，现有的RS数据集通常通过自动图像描述或传统元数据处理流程构建，因而往往存在不完整、易出错或语义—

无关的描述。因此，相当一部分图文对仅部分对齐，甚至完全不匹配，从而引入噪声监督，误导那些在完美对应这一假设下训练的模型。为了在此类不完善的标注下维持稳健的检索能力，不仅必须抑制伪匹配对的有害影响，还应

*通讯作者 版权所有 © 2026年，人工智能促进协会 (www.aaai.org)。保留所有权利。

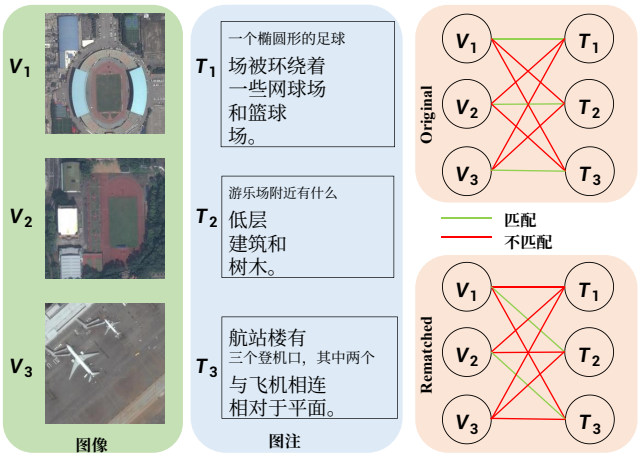


图 1：一个简单的例子展示了我们的核心见解：伪匹配对（PMP）——存在部分对齐或对齐错误的图文样本——并非只是噪声，往往还包含潜在的语义线索。PMPGuard 并未将其丢弃，而是利用这些线索重新匹配语义相关的样本对（绿色连线），并排斥不相关的样本对（红色连线），从而有效地将噪声监督转化为有用的对齐信号。

从中提炼潜在的有用线索（Zhang 等人（2024））。

大多数现有方法（Yu 和 Ji，2022；Wang 等，2022；Chen 等，2022）依赖标注完善的图文对来学习有效的联合表征。然而，在遥感领域，收集高质量且精确对齐的图像–标题对成本极高，而且往往难以实现。因此，许多数据集采用网络爬取或自动生成的描述构建，这不可避免地引入了伪匹配对（PMP）——即图像与图注仅部分对齐或在语义上不一致的样本对。为缓解这一问题，近期研究尝试降低伪匹配对的影响，或放宽基于排序的目标函数中的损失边界。然而，这些方法往往未能充分利用不匹配样本，只将其视为需要抑制的噪声，而非可能蕴含信息、值得挖掘的信号。

如图1所示，一个简单的示例阐释了我们的

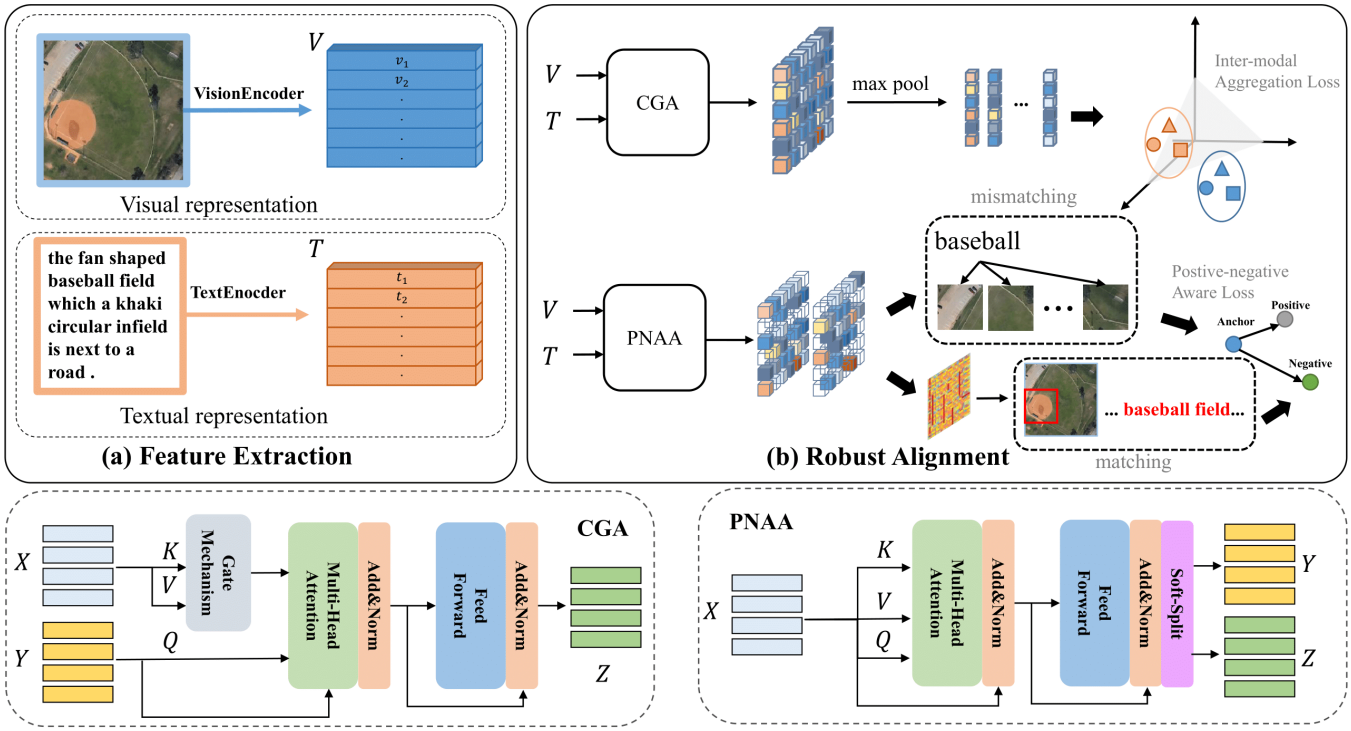


图2：PMPGuard 概览：视觉和文本编码器提取特征，随后交叉门控注意力（CGA）抑制未匹配线索，正负感知注意力（PNAA）则对比可靠与不可靠的图文对；二者共同优化，以减少伪匹配对并加强跨模态对齐。

核心思想：未配对样本之间潜在的语义相似性，使我们能够从不匹配对中提取有价值的知识。在遥感图像-文本检索中，此类伪匹配对十分常见，图文样本具有部分语义相关性，但并未实现完美对齐。这带来了两项主要挑战：(1) 区分真正不匹配的元素与部分相关的元素；(2) 在不受噪声信号误导的情况下，利用不匹配对中的有用信息。为应对这些问题，我们提出了**PMPGuard**，一种通过明确识别和利用伪匹配对中的潜在语义线索来增强跨模态对齐的新型框架。PMPGuard 并不将不匹配对简单地视为噪声而丢弃，而是重新匹配语义相关的跨模态样本（绿色连接），并排斥不相关样本（红色连接），从而带来更稳健的检索性能。我们的主要贡献总结如下：

- 我们提出了一种新颖的检索框架，**PMPGuard**，专为解决由**伪匹配对（PMP）**造成的语义错位和噪声监督而设计。伪匹配对是指仅部分对齐或对齐错误的图文对，在大规模遥感数据集中十分常见。**PMPGuard**并非只是抑制伪匹配对，而是识别并利用它们来加强跨模态对齐。
- 我们提出了一种**跨门控注意力（CGA）**模块，可自适应地调节跨模态特征交换。CGA 会选择性地促进语义一致的内容流动，同时在训练过程中滤除模态特定噪声和由伪匹配对（PMP）引起的不匹配项。
- 我们设计了**正负感知注意力（PNAA）**模块，用于明确区分图文对中的对齐区域和未对齐区域。PNAA 通过双分支结构对正负线索进行建模，从而增强对伪匹配对所带来噪声监督的鲁棒性。

•在三个公开遥感基准数据集（即 **RSICD**、**RSITMD** 和 **RS5M**）上进行的大量实验表明，**PMPGuard** 始终能够达到**最先进水平（SOTA）**。值得注意的是，在错配率较高的情况下，其优势更加明显，证实了其在存在伪匹配对时的稳健性和泛化能力。

相关工作

遥感图文检索

遥感图文检索旨在将航空影像与自然语言进行对齐，从而支持土地覆盖分析和灾害响应等应用。近年来，研究重点已从传统的CNN-RNN框架转向具有更细粒度对齐能力的基于Transformer的模型。例如，PIR（Pan、Ma和Bai，2023a）将先验知识融入检索过程，而SWAN（Pan、Ma和Bai，2023b）引入场景感知聚合，以减少语义不一致。

DOVE（Ma、Pan和Bai，2024）利用视觉语义嵌入增强方向性对齐。这些方法通常假设训练对具有良好的对齐关系，这限制了它们在噪声真实世界数据集上的鲁棒性。为应对不完善监督，近期的BiCro（Yang等，2023）和DEL（Feng等，2024）等工作提出了噪声感知训练策略。然而，它们大多只是抑制不匹配项，而没有加以利用。我们的方法则通过门控注意力和正负样本感知主动挖掘不匹配对，从而在存在噪声对应关系的情况下实现更稳健的对齐。

抗不匹配检索模型

抗不匹配检索致力于在训练数据包含噪声或伪匹配图文对时，学习可靠的跨模态对齐。这对于遥感数据集尤为重要，因为大规模标注通常存在噪声或采用弱监督方式。DEL（Feng 等，2024年）引入证据学习来估计不确定性，并降低错误图文对的影响。BiCro（Yang 等，2023）通过强制双向相似度一致性来纠正噪声对应关系。L2RM（Han 等，2024年）采用联合重新对齐策略，在训练过程中显式地对不匹配对进行重新匹配。尽管这些方法颇为有效，但大多数现有方法都将不匹配对视为需要抑制的噪声。相比之下，我们的方法不仅能够识别不匹配，还能通过门控注意力和判别式挖掘利用其中的潜在语义关系，从而实现更加稳健且信息更丰富的对齐（Cheng 等，2021年）。

方法

问题定义

不失一般性，我们以视觉文本检索为例，形式化定义跨模态检索中的伪匹配对（PMP）问题。设训练集为

$$\mathcal{D} = \{(V_i, T_i, m_i)\}_{i=1}^{N},$$

其中，(V_i，T_i)表示一个视觉文本对，$m_i \in \{0,1\}$表示该对在语义上是否匹配。

跨模态检索的核心在于衡量异构模态之间的相似度。现有方法首先通过模态专属编码器 f_v 和 f_t，分别将视觉输入和文本输入投影到共享嵌入空间，然后计算一对(V_i，T_j)的相似度，公式如下：

$$S_{ij} = g\big(f_v(V_i),\, f_t(T_j)\big),$$

其中，g 是参数映射或非参数映射。

大多数以往方法都默认每个标注为 $m_i = 1$ 的配对确实匹配。然而，现实中的遥感数据集通常通过自动化图像描述生成流程、众包或历史元数据整理而成，其中往往存在相当比例的不匹配视觉文本对，却被错误地标注为正样本。这些伪匹配对（PMP）会传播噪声监督信号，并损害模型的鲁棒性。因此，我们的目标是检测并抑制伪匹配对的影响，从而实现面向遥感数据的鲁棒跨模态检索。

交叉门控注意力机制

为解决伪匹配对（如引言中所述）导致的噪声特征交互问题，我们的交叉门控注意力机制会动态调节信息流。具体而言，它解决两个子问题：（1）抑制可能由不匹配对引发的无关特征；（2）保留可能有用的跨模态交互。

给定文本特征序列 $\mathcal{U} = \{u_i\}_{i=1}^{m}$ 和一组图像区域特征 $\mathcal{V} = \{v_j\}_{j=1}^{n}$，目标是通过相互注意力和门控操作，计算模态感知表示。

交叉注意力。首先计算每个词 u_i 与图像区域 v_j 之间的注意力分数：

$$a_{ij} = \frac{u_i^\top W_a v_j}{\|u_i\|\|v_j\|}, \tag{1}$$

其中，W_a 是一个可学习投影矩阵。随后使用soft max对注意力权重进行归一化：

$$\alpha_{ij} = \frac{\exp(a_{ij})}{\sum_{j'=1}^{n}\exp(a_{ij'})}, \beta_{ji} = \frac{\exp(a_{ij})}{\sum_{i'=1}^{m}\exp(a_{i'j})}. \tag{2}$$

利用这些注意力权重，我们生成跨模态注意力特征：

$$\tilde{v}_i = \sum_{j=1}^{n}\alpha_{ij}v_j, \tilde{u}_j = \sum_{i=1}^{m}\beta_{ji}u_i. \tag{3}$$

门控机制。为了抑制无关或噪声特征并突出重要交互，我们引入门控函数，以控制从每种模态中保留的信息量。门控表示的计算方式如下：

$$g_i^u = \sigma(W_g^u[\mathbf{u}_i;\tilde{\mathbf{v}}_i] + b_g^u), g_j^v = \sigma(W_g^v[\mathbf{v}_j;\tilde{\mathbf{u}}_j] + b_g^v), \tag{4}$$

其中，$[;\cdot]$ 表示拼接，σ 表示 sigmoid 激活函数，W_g^u, W_g^v和 b_g^u, b_g^v是可学习参数。$\mathbf{u}_i$ 和 $\mathbf{v}j$ 是原始的模态特定嵌入，$\tilde{}\mathbf{v}_i$和 $\tilde{}\mathbf{u}j$ 是跨模态上下文，而 g_i^u，$g_j^v \in [0,1]^d$ 是学得的门控向量。

最终的交叉门控特征为：

$$\hat{\mathbf{u}}_i = g_i^u \odot \mathbf{u}_i + (1-g_i^u)\odot\tilde{\mathbf{v}}_i, \hat{\mathbf{v}}_j = g_j^v \odot \mathbf{v}_j + (1-g_j^v)\odot\tilde{\mathbf{u}}_j, \tag{5}$$

其中，$\odot$ 表示逐元素乘法。该机制使模型能够选择性地融合模态内特征和模态间特征，从而为图文匹配生成更具表现力的表示。

模态间聚合损失 为对齐经过交叉门控注意力（CGA）精炼后的视觉和语言表示，我们引入模态间聚合损失（IA）。令 $\hat{}\mathbf{u}_i$ 和 $\hat{}\mathbf{v}j$ 分别表示门控后的视觉特征和文本特征。该损失定义为带温度参数 τ 的 InfoNCE 目标：

$$\mathcal{L}_{\text{IA}} = -\sum_{(i,j)\in\mathcal{P}} \log \frac{\exp\left(\frac{\hat{\mathbf{u}}_i^\top \hat{\mathbf{v}}_j}{\tau}\right)}{\sum_k \exp\left(\frac{\hat{\mathbf{u}}_i^\top \hat{\mathbf{v}}_k}{\tau}\right) + \sum_k \exp\left(\frac{\hat{\mathbf{u}}_k^\top \hat{\mathbf{v}}_j}{\tau}\right)}, \tag{6}$$

其中，$\mathcal{P}$ 表示语义匹配对的集合。通过最大化配对样本之间的一致性，同时拉开非配对样本之间的距离，该损失增强了跨模态对齐的鲁棒性，并减轻了伪匹配对带来的不利影响。

正负感知注意力

为应对伪匹配对兼具误导性和潜在实用性的双重特性，PNAA 通过以下方式显式建模正负信号：(1) 负分支识别并抑制真正不匹配的片段（式 12–13）；(2) 正分支从部分匹配区域中提取有用的语义线索（式 14–15）。这一双重策略直接应对了问题陈述中提出的如何利用不完美训练对这一挑战。

捕捉伪匹配对为区分匹配和不匹配的图文对，GeoRSCLIP（Zhang 等人（2024））将它们的相似度分布建模为高斯函数：

$$f_k^+(s) = \frac{1}{\sigma_k^+\sqrt{2\pi}} \exp\left(-\frac{(s-\mu_k^+)^2}{2(\sigma_k^+)^2}\right), \tag{7}$$

$$f_k^-(s) = \frac{1}{\sigma_k^-\sqrt{2\pi}} \exp\left(-\frac{(s-\mu_k^-)^2}{2(\sigma_k^-)^2}\right), \tag{8}$$

其中 μ_k^+，σ_k^+ 和 μ_k^-，σ_k^- 分别表示匹配对和不匹配对的相似度分数的均值和标准差。

为自动分离这两个分布，GeoRSCLIP 通过最小化两个分布的加权重叠来学习决策边界 t_k：

$$t_k = \arg\min_{t\geq 0} \left[\alpha \int_t^{+\infty} f_k^-(s)\,ds + \int_{-\infty}^{t} f_k^+(s)\,ds\right], \tag{9}$$

其中，α 是用于平衡假阳性和假阴性的惩罚参数。随后，利用学习到的边界 t_k 在注意力计算过程中分离正片段和负片段，从而在噪声监督下实现更稳健的对齐。

负分支。对于每个词语 u_i，我们计算其与所有图像区域之间的最大相似度：

$$s_i = \max_j \left(\cos(u_i, v_j) - t_k\right), \tag{10}$$

并应用负掩码来抑制匹配的词语：

$$s_i^{\text{neg}} = s_i \cdot \text{Mask}_{\text{neg}}(s_i), \tag{11}$$

其中，如果 $\text{Mask}^{\text{neg}}(s_i) = 1$，则为 $s_i < 0$，否则为 0。

正分支。我们计算跨模态注意力权重：

$$w_{ij}^{\text{inter}} = \text{softmax}\left(\text{Mask}_{\text{pos}}(s_{ij} - t_k)\right), \tag{12}$$

其中，当 $s_i > 0$ 时，$\text{Mask}^{\text{neg}}(s_i) = 1$，否则为 0。然后聚合匹配的图像特征，得到正分数：

$$s_i^{\text{pos}} = \cos(u_i, \hat{v}_i) + \sum_j w_{ij}^{\text{relev}} s_{ij}, \tag{13}$$

其中 $\hat{v}_i = \sum_j w_{ij}^{\text{inter}} v_j$，$w_{ij}^{\text{relev}}$是基于相关性的注意力权重。

正负感知损失 最终的相似度分数计算如下：

$$S(\mathcal{V},\mathcal{U}) = \frac{1}{m}\sum_{i=1}^{m}\left(s_i^{\text{pos}} + s_i^{\text{neg}}\right). \tag{14}$$

我们采用双向三元组排序损失进行训练：

$$\begin{aligned}\mathcal{L}_{\mathcal{PA}} = &\sum_{(\mathcal{U},\mathcal{V})} [\gamma - S(\mathcal{U},\mathcal{V}) + S(\mathcal{U},\mathcal{V}')]_+ \\ &+ [\gamma - S(\mathcal{U},\mathcal{V}) + S(\mathcal{U}',\mathcal{V})]_+,\end{aligned} \tag{15}$$

其中 γ 是一个间隔超参数，$[\cdot]_+$ 表示合页损失。

训练目标

最终的训练目标是联合最小化模态间聚合损失和正负感知损失：

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{IA}} + \lambda\,\mathcal{L}_{\text{PA}}, \tag{16}$$

其中 $\lambda > 0$ 平衡这两项。最小化 $\mathcal{L}_{\text{total}}$能够通过基于 InfoNCE 的 $\mathcal{L}_{\text{IA}}$ 同时对齐跨模态表示，并通过基于间隔的 $\mathcal{L}_{\text{PA}}$抑制噪声或不匹配对的影响，从而在不完善监督下实现稳健检索。

实验

数据集

RSICD RSICD（Lu 等人，2018）是专为遥感图像描述任务打造的基准数据集，包含 10,921 张地理分布多样的图像，取自 Google Earth 和街景；每张图像均配有五条简洁的英文描述。该数据集涵盖城市、乡村、农业、山区和水域场景，所有图像均已统一调整为224×224 像素，以便进行一致的处理。

RSITMD RSITMD（Yuan 等人，2022年）是一个用于跨模态遥感图像检索的细粒度基准。它包含 4,743 幅图像，与 23,715 条多样且低冗余的图注配对，并为每幅图像提供 1–5 个精准关键词，从而形成更丰富的对象级描述。RSITMD 旨在降低类内相似度并提升类间多样性，相较于现有数据集，能够提供更优的检索准确率。

MRate	方法	RSICD							RSITMD						
		句子检索			图像检索			mR	句子检索			图像检索			mR
		R1	R5	R10	R1	R5	R10		R1	R5	R10	R1	R5	R10	
0	L2RM	5.01	13.19	21.84	4.32	15.21	26.30	14.15	7.37	23.07	35.73	7.42	29.76	49.77	25.45
	HarMA-Vit	1.49	3.83	7.53	1.15	3.47	6.80	3.97	2.89	4.82	9.67	1.61	4.61	9.63	5.57
	PIR	8.61	23.57	35.40	5.49	21.84	38.04	21.99	11.14	33.14	45.21	8.60	33.02	56.43	30.99
	SWAN	9.82	24.61	36.25	6.70	22.51	38.67	23.48	11.91	34.29	46.16	9.50	33.66	57.14	32.63
	DOVE	10.54	25.31	38.27	7.93	23.54	40.13	23.79	13.51	35.25	48.12	10.21	34.52	58.45	33.29
	PMPGuard	**12.56**	**28.20**	**39.37**	**9.76**	**25.55**	**42.49**	**26.21**	**14.94**	**37.39**	**49.11**	**12.58**	**37.06**	**59.01**	**34.95**
0.2	L2RM	4.03	12.35	21.04	3.59	14.49	25.51	13.50	6.64	22.35	34.96	6.86	28.94	48.94	24.78
	HarMA-Vit	0.82	3.29	6.50	0.49	2.71	5.65	3.24	2.21	3.98	8.63	0.88	3.76	8.89	4.73
	PIR	7.78	22.69	34.40	4.85	20.70	36.94	21.23	10.4	32.08	44.25	7.65	32.08	55.31	30.29
	SWAN	8.95	23.72	35.07	5.85	21.66	37.91	22.66	11.25	33.25	45.06	8.85	32.85	56.04	31.87
	DOVE	9.88	24.25	37.34	7.04	22.82	39.43	23.11	12.57	34.55	47.01	9.61	33.66	57.46	32.48
	PMPGuard	**11.90**	**27.26**	**38.20**	**9.01**	**24.42**	**41.67**	**25.42**	**14.06**	**36.64**	**48.46**	**11.87**	**36.09**	**57.88**	**34.07**
0.4	L2RM	2.84	9.06	17.11	2.73	11.40	20.55	10.61	5.09	17.92	30.09	4.47	20.13	35.49	18.86
	HarMA-Vit	1.01	3.20	6.86	0.59	1.99	4.96	3.10	1.77	5.31	7.96	0.93	3.98	7.30	4.54
	PIR	6.22	17.38	29.37	4.41	18.50	32.81	18.12	9.51	26.33	40.27	7.08	29.56	51.59	27.39
	SWAN	7.44	22.31	33.60	4.41	20.29	36.50	21.19	9.82	31.77	43.69	7.33	31.36	54.47	30.26
	DOVE	8.35	22.95	35.96	5.48	21.31	37.86	21.77	11.06	33.17	45.60	8.21	32.18	56.01	30.91
	PMPGuard	**10.38**	**25.77**	**36.90**	**7.44**	**22.92**	**40.18**	**24.03**	**12.61**	**35.29**	**47.25**	**10.41**	**34.63**	**56.62**	**32.58**
0.6	L2RM	0.88	2.21	3.32	0.35	2.35	4.56	2.28	3.54	15.04	25.44	4.20	18.10	29.47	15.97
	HarMA-Vit	0.64	2.84	5.40	0.60	2.51	4.56	2.76	1.11	5.75	7.74	0.40	3.50	6.33	4.14
	PIR	4.67	13.72	25.07	2.95	14.82	26.55	14.63	9.96	27.88	38.72	6.46	26.28	48.36	26.28
	SWAN	6.59	21.61	32.44	3.63	19.51	35.58	20.34	8.94	31.21	43.22	6.59	30.80	53.65	29.31
	DOVE	7.49	22.30	34.80	4.64	20.48	36.86	20.97	10.28	32.66	44.93	7.53	31.51	55.11	30.34
	PMPGuard	**9.51**	**24.94**	**35.70**	**6.66**	**22.05**	**39.23**	**23.03**	**11.71**	**34.74**	**46.61**	**9.58**	**33.87**	**55.83**	**31.54**
0.8	L2RM	0.22	1.55	3.32	0.58	2.35	4.56	2.49	3.54	11.28	17.92	1.86	7.88	14.56	9.51
	HarMA-Vit	0.27	1.92	3.66	0.40	1.70	3.17	1.85	0.88	3.32	4.87	0.35	2.04	4.96	2.74
	PIR	3.66	11.16	21.04	2.34	11.4	21.67	11.88	4.65	17.04	27.43	3.63	15.44	28.98	16.19
	SWAN	4.61	19.64	31.30	1.69	17.55	33.71	18.53	7.95	29.97	41.88	4.55	28.91	51.86	27.41
	DOVE	5.57	20.24	33.57	2.75	18.56	35.02	19.01	8.38	30.91	43.94	5.53	29.65	53.97	28.68
	PMPGuard	**7.48**	**22.80**	**34.40**	**4.74**	**20.15**	**37.30**	**21.12**	**10.01**	**32.90**	**44.97**	**7.48**	**31.72**	**54.77**	**29.64**

表1：RSICD 和 RSITMD 上不同 MRate 下的图文检索性能。

RS5M RS5M（Zhang 等人（2024））是一个包含五百万对数据的遥感视觉语言语料库，其图注通过从开放平台自动抓取获得。庞大的规模也带来了大量伪匹配对（PMP）：简短、模板化或完全不相关的文本，常常与农业、基础设施或土地利用图像一同出现，从而造成嘈杂、薄弱且地理分布偏斜的对齐。质量评分子集显示，最多有 30 % 的数据对存在显著不匹配，这使 RS5M 成为研究现实、不完美监督条件下检索鲁棒性的理想测试平台。

实现细节

本研究的所有实验均使用两块 NVIDIA RTX A6000 GPU 进行。为确保不同图像尺寸的数据集之间的一致性，我们将图像统一调整为 224×224 像素后输入网络。我们采用了旋转和翻转等数据增强技术，以提升模型鲁棒性。词向量的表示维度为 768，而图像和文本的嵌入空间维度为 512。参数 τ 初始化为 0.07，并在计算对比损失时进行学习；计算三元组损失时，间隔设为 0.5。网络使用 AdamW 优化器训练 20 个训练轮次，同时结合对比损失和三元组损失，每块 GPU 的批量大小为 128。为确保实验结果的鲁棒性，每组实验均进行五次，并记录平均结果。

基线方法

我们将 PMPGuard 与五种最先进的跨模态检索方法进行比较，其中包括五种通用方法：PIR（Pan、Ma 和 Bai，2023a）、SWAN（Pan、Ma 和 Bai，2023b）、HarMA（Huang，2024）、DOVE（Ma、Pan 和 Bai，2024），此外还有 L2RM（Han 等人，2024）。采用 GeoRSCLIP（Zhang 等人（2024））作为骨干网络的 L2RM 和 HarMa 被视为大参数模型。

主要结果

伪匹配训练数据的构建 为评估在不完美监督下的鲁棒性，我们模拟

MRate	方法	RSICD							RSITMD						
		句子检索			图像检索			mR	句子检索			图像检索			mR
		R1	R5	R10	R1	R5	R10		R1	R5	R10	R1	R5	R10	
0	L2RM-Geo	18.84	42.30	53.65	14.86	38.60	55.55	37.42	27.98	52.80	62.83	22.95	55.20	72.34	48.71
	HarMA-Geo	19.51	43.72	55.68	15.41	40.06	57.73	38.90	29.04	54.56	65.03	23.82	56.95	74.61	50.39
	PMPGuard-Geo	**20.21**	**45.05**	**57.28**	**15.93**	**41.30**	**59.39**	**40.23**	**30.12**	**56.21**	**67.00**	**24.55**	**58.75**	**76.84**	**52.01**
0.2	L2RM-Geo	17.75	40.65	52.55	13.68	37.50	53.82	36.10	27.00	51.05	61.38	21.85	53.40	70.40	47.25
	HarMA-Geo	18.39	42.36	54.53	14.29	38.96	56.25	37.46	28.10	53.10	63.72	22.70	55.62	73.58	49.47
	PMPGuard-Geo	**19.02**	**43.85**	**56.30**	**14.85**	**40.32**	**58.11**	**38.75**	**29.15**	**54.86**	**66.10**	**23.50**	**57.48**	**76.00**	**51.10**
0.4	L2RM-Geo	17.44	38.76	51.80	12.92	36.85	52.28	34.95	26.75	50.10	61.92	21.78	53.25	70.50	47.30
	HarMA-Geo	18.12	40.35	53.98	13.50	38.19	54.47	36.43	27.88	52.21	64.38	22.74	55.53	73.72	49.41
	PMPGuard-Geo	**18.75**	**41.96**	**56.18**	**14.02**	**39.50**	**56.86**	**37.70**	**28.95**	**54.10**	**66.90**	**23.45**	**57.30**	**76.05**	**51.12**
0.6	L2RM-Geo	16.45	38.45	51.47	11.75	33.90	50.90	33.72	25.40	48.70	61.80	19.85	51.20	67.25	45.65
	HarMA-Geo	17.11	40.07	53.71	12.28	35.24	53.14	35.26	26.55	50.88	64.16	20.75	53.50	70.31	47.69
	PMPGuard-Geo	**17.68**	**41.45**	**55.60**	**12.78**	**36.55**	**55.35**	**36.50**	**27.58**	**52.80**	**66.50**	**21.50**	**55.25**	**73.00**	**49.35**
0.8	L2RM-Geo	17.54	36.33	48.33	11.56	34.27	50.50	33.06	25.01	47.08	60.33	19.50	49.65	66.38	44.66
	HarMA-Geo	18.30	38.24	50.87	12.17	36.07	53.16	34.80	26.33	49.56	63.50	20.53	52.26	69.87	47.01
	PMPGuard-Geo	**19.07**	**40.15**	**53.41**	**12.78**	**37.87**	**55.82**	**36.54**	**27.65**	**52.04**	**66.68**	**21.56**	**54.87**	**73.36**	**49.36**

表 2：在不同误匹配率（MRate）下，以大型模型 GeoRSCLIP 为骨干网络时，其在 RSICD 和 RSITMD 上的图文检索性能。

方法	RS5M						
	句子检索			图像检索			mR
	R1	R5	R10	R1	R5	R10	
L2RM-Geo	13.58	34.42	45.21	9.84	31.26	46.90	30.42
HarMA-Geo	14.84	37.13	48.39	10.96	33.74	50.12	32.78
PMPGuard-Geo	**16.12**	**39.80**	**51.60**	**12.10**	**36.05**	**53.43**	**35.21**

表 3：RS5M 上的图文检索性能。

通过错配可控的伪匹配训练集来模拟真实噪声。对于干净数据集 $\mathcal{D} = (V_i,\ T_i)_{i=1}^{N}$，我们通过将 x% 的配对中的 T_i 替换为随机的 T_j $(j \neq i)$，生成 $\mathcal{D}_x$。为确保语义上确实不匹配，我们使用预训练的 **GeoRSCLIP**模型计算相似度分数，并仅保留分数低于阈值 τ的替换项。此筛选可避免简单或语义相似的替换，从而提高噪声的真实性。我们构建了多个错配率不同的变体（$x = 0$%，20%.....，80%），以在不同噪声水平下对检索模型进行基准测试。

合成伪匹配对上的结果表 1 展示了我们在 RSICD 和 RSITMD 上的实验结果，其中模型采用 Swin Transformer（Liu 等，2021）作为图像骨干网络，采用 BERT（Devlin 等，2019）作为文本骨干网络。从结果可以看出，与其他最先进的方法相比，PMPGuard 在所有指标上均取得了最佳表现，展现出强大的鲁棒跨模态检索能力。此外，当错配率较高（如 0.6 和 0.8）时，PMPGuard 带来的提升更加显著，验证了错配对挖掘对于增强检索鲁棒性的有效性。

基于大型预训练骨干网络的评估结果表 2 展示了不同方法在 RSICD 和 RSITMD 数据集上面对不同错配率（MRate）时的表现。总体而言，我们提出的方法 **PMPGuard-Geo** 在所有设置和评估指标上均持续优于基线方法（L2RM-Geo 和 HarMA-Geo）。在 RSICD 上，PMPGuard-Geo 在每个错配水平下都取得了最高的平均召回率，从 37.42（L2RM-Geo）和 38.90（HarMA-Geo）显著提升至 MRate = 0 时的 40.23。即使错配率升至 0.8，PMPGuard-Geo 仍保持出色表现（平均召回率 = 36.54），展现出对伪匹配对的卓越鲁棒性。在 RSITMD 上，PMPGuard-Geo 也表现出类似优势，持续取得最佳的 R@1、R@5、R@10 和平均召回率。例如，在 MRate= 0.8 时，PMPGuard-Geo 的平均召回率达到 49.36，显著优于 L2RM-Geo（44.66）和 HarMA-Geo（47.01）。这些结果验证了 PMPGuard-Geo 不仅能更有效地捕捉细粒度跨模态语义，而且对噪声监督也更具鲁棒性，因此非常适用于真实的遥感图文检索任务。

真实伪匹配对上的结果 我们的主要发现凸显了 PMPGuard-Geo 在解决遥感图像-文本检索中的伪匹配对（PMPs）方面的有效性。如表 3 所示，PMPGuard-Geo 实现了同类最佳的检索性能，在句子检索和图像检索的 R1 得分上均取得显著提升，展现出较高的精确率，并能够通过交叉门控注意力（CGA）抑制噪声。R@10 的持续提升进一步验证了正负样本感知注意力（PNAA）的有效性，该机制能够有效区分潜在语义线索与噪声。特别是，PMPGuard-Geo 将平均召回率（mR）提升了 4.79%，证明了其在不同检索方向和阈值下的稳健性。此外，在 RS5M 数据集上相较于合成噪声基线的性能差距，

MRate	方法	RSICD							RSITMD						
		句子检索			图像检索			mR	句子检索			图像检索			mR
		R1	R5	R10	R1	R5	R10		R1	R5	R10	R1	R5	R10	
0	不含 CGA	18.32	41.65	52.70	14.22	37.45	54.80	36.52	27.15	51.90	61.30	22.10	54.30	71.40	47.53
	不含 PNAA	19.35	43.20	55.12	15.12	39.65	56.90	38.56	28.45	53.95	64.20	23.35	56.40	73.95	49.38
	full	**20.21**	**45.05**	**57.28**	**15.93**	**41.30**	**59.39**	**40.23**	**30.12**	**56.21**	**67.00**	**24.55**	**58.75**	**76.84**	**52.01**
0.2	不含 CGA	17.25	39.45	50.83	13.10	36.40	52.70	34.96	26.10	49.90	60.20	20.70	52.30	69.20	46.57
	不含 PNAA	18.15	41.30	53.20	13.85	38.05	54.85	36.90	27.30	52.40	62.90	22.05	54.90	72.60	48.69
	full	**19.02**	**43.85**	**56.30**	**14.85**	**40.32**	**58.11**	**38.75**	**29.15**	**54.86**	**66.10**	**23.50**	**57.48**	**76.00**	**51.10**
0.4	不含 CGA	16.85	37.80	49.65	12.50	35.50	51.40	33.45	25.60	48.80	59.80	20.35	50.80	67.90	45.54
	不使用 PNAA	17.76	39.90	52.30	13.20	37.10	53.20	35.91	26.65	51.30	63.30	21.45	53.30	71.45	48.24
	full	**18.75**	**41.96**	**56.18**	**14.02**	**39.50**	**56.86**	**37.70**	**28.95**	**54.10**	**66.90**	**23.45**	**57.30**	**76.05**	**51.12**
0.6	不含 CGA	15.90	36.40	48.75	11.30	32.80	49.85	32.50	24.15	47.10	59.60	18.80	49.85	65.20	44.12
	不使用 PNAA	16.68	38.20	51.10	12.05	34.40	51.70	34.69	25.65	49.85	62.30	20.00	52.40	68.95	46.52
	full	**17.68**	**41.45**	**55.60**	**12.78**	**36.55**	**55.35**	**36.50**	**27.58**	**52.80**	**66.50**	**21.50**	**55.25**	**73.00**	**49.35**
0.8	不含 CGA	15.40	34.50	46.00	10.75	31.40	48.10	31.36	23.45	45.85	58.00	18.10	48.05	63.30	42.79
	不使用 PNAA	16.88	36.80	49.45	11.80	33.85	50.65	33.57	24.95	48.60	61.30	19.30	51.20	66.75	45.02
	full	**19.07**	**40.15**	**53.41**	**12.78**	**37.87**	**55.82**	**36.54**	**27.65**	**52.04**	**66.68**	**21.56**	**54.87**	**73.36**	**49.36**

表 4：PMPGuard 模型在 RSICD 和 RSITMD 上不同错配率（MRate）下的消融实验。

进一步展现了 PMPGuard 独特的真实世界伪匹配对挑战处理能力，包括自然对齐不一致、地理变化以及大规模噪声监督。这些结果证实，显式建模正负样本关系并自适应地门控跨模态信号，能够将伪匹配对从学习障碍转变为有价值的监督信息。

消融实验

如表 4 所示，我们在 RSICD 和 RSITMD 数据集上针对不同 MRate 进行了消融研究，以评估所提出的 P MPGuard-Geo 框架中各组件的有效性。我们考察了移除 CGA 或 PNAA 模块的变体（分别记为“w/o CGA”和“w/o PNAA”），并将其与完整模型进行比较。移除 CGA 或 PNAA 中的任一模块，都会导致模型在所有指标和错配水平下的性能持续下降。尤其是，移除 PNAA 会导致模型在较高错配率下性能更明显地下降：例如，MRate =0.8 时，RSITMD 上的平均召回率（mR）从 49.36 降至 45.02，表明 PNAA 在处理噪声监督方面效果显著。省略 **CGA** 会普遍降低召回率，尤其是在句子检索中，这表明它有助于增强跨模态对齐。总体而言，**CGA** 和 **PNAA** 发挥互补作用：CGA 强化语义对应关系，而 PNAA 提高模型对噪声或弱对齐对的鲁棒性，二者共同使 PMPGuard-Geo 能够保持

可视化实验

图 3 展示了 PMPGuard 修正遥感（RS）图像-文本不匹配对的能力。图中呈现了三个最初错误的图注-图像匹配案例，PMPGuard 成功为每幅图像重新匹配了更恰当的描述——绿色标示正确匹配，红色表示原始不匹配项。这体现了 PMPGuard



图3：PMPGuard 重新匹配不匹配的遥感图文对；绿色和红色文字分别表示正确和错误的匹配。

在增强遥感图像-文本语义对齐和跨模态检索准确率方面的有效性。这项关键能力适用于需要精准理解图像的实际应用。

结论

我们提出**PMPGuard**，一种用于应对伪匹配对（P MP）的鲁棒遥感（RS）图文检索框架，具有两项核心创新：**交叉门控注意力（CGA）**（动态跨模态特征调制）和**正负感知注意力（PNAA）**（判别式对齐学习）。在 RSICD、RSITMD 和 RS5M 上进行的大量实验表明，该方法在 0–0.8 的错配率范围内均表现出色，取得了最先进的结果，并能抵御噪声对应关系，为现实世界中存在不完美对齐的检索场景提供了有理论依据的解决方案，也为鲁棒的跨模态遥感学习奠定了基础。

致谢

本研究部分得到了浙江省自然科学基金项目（编号：LY24F020020 和 LRG25F020002）以及国家自然科学基金项目（编号：U20A20196 和 62302453）的资助。

参考文献

Chen, J.; Huang, H.; Peng, J.; Zhu, J.; Chen, L.; Tao, C.; and Li, H.2022. 面向遥感场景分类的上下文信息保留架构学习。*IEEE*地球科学与遥感汇刊, 60: 1–14.

Chen, Y.; Du, C.; Zi, Y.; Xiong,S.;andLu, X. 2024a. 面向遥感视听跨模态检索的尺度感知自适应细化与交叉交互。*IEEE*地球科学与遥感汇刊。

Chen, Y.; Huang, J.; Li, X.; Xiong, S.; and Lu, X. 2023. 面向遥感图文检索的多尺度显著性对齐学习。*IEEE*地球科学与遥感汇刊, 62: 1–13. Chen, Y.; Huang, J.; Xiong, S.; and Lu, X. 2024b. 将多子空间联合学习与多层级引导相结合，用于遥感图像跨模态检索。*IEEE*地球科学与遥感汇刊, 62: 1–17. Cheng, Q.; Zhou, Y.; Fu, P.; Xu, Y.; and Zhang, L. 2021. 面向遥感跨模态图文检索的深度语义对齐网络。*IEEE*《应用地球观测与遥感精选主题期刊》, 14: 4284–4297. Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT：用于语言理解的深度双向 Transformer 预训练。载于 Burstein, J.; Doran, C.; and Solorio, T. 编，*2019*年北美计算语言学协会分会会议论文集：人类语言技术，*NAACL-HLT 2019*，美国明尼阿波利斯，*2019*年*6*月*2*日至*7*日，第*1*卷（长篇与短篇论文）, 4171–4186. 计算语言学协会。El man, J. L. 1990. 时间结构发现。认知科学, 14(2): 179–211. Faes, M. G.; Broggi, M.; Spanos, P. D.; and Beer, M. 2022. 阐明可微自相关函数的吸引性特征：修正指数核研究。《概率工程力学》, 69: 103269. Feng, Z.; Zeng, Z.; Guo, C.; Li, Z.; and Hu, L. 2024. 利用三分区从噪声对应关系中学习跨模态匹配。*IEEE*多媒体汇刊, 26: 3884–3896. Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. 通过学习不变映射进行降维。载于*2006*年*IEEE*计算机学会计算机视觉与模式识别会议（*CVPR'06*）, 第2卷, 1735–1742. IEEE。Han, H.; Zheng, Q.; Dai, G.; Luo, M.; and Wang, J. 2024. 学习重新匹配不匹配对，以实现稳健的跨模态检索。载于*IEEE/CVF*计算机会议视觉与模式识别，*CVPR 2024*，美国华盛顿州西雅图，*2024*年*6*月*16*日至*22*日，26669–26678。IEEE。He, Y.; Xu, X.; Chen, H.; Li, J.; and Pu, F. 2024. 面向遥感图像-文本检索的视觉全局显著性引导网络。*I EEE*地球科学与遥感汇刊。Hu, P.; Huang, Z.; Peng, D.; Wang, X.; and Peng, X. 2023. 基于部分不匹配对的跨模态检索。*IEEE* 模式分析与机器智能汇刊，45(8): 9595–9610。Huan, R.; Zhong, G.; Chen, P.; and Liang, R. 2024a. TriSAT：面向多模态情感分析的三模态表征学习。*IEEE ACM*音频、语音与语言处理汇刊, 32: 4105–4120。Huan, R.; Zhong, G.; Chen, P.; and Liang, R. 2024b. UniMF：针对模态缺失和未对齐多模态序列的多模态情感分析统一框架。*IEEE*多媒体汇刊, 26: 5753–5768。Huan, R.; Zhong, G.; Chen, P.; and Liang, R. 2025. MulDeF：用于鲁棒多模态情感分析的模型无关去偏框架。*IEEE*多媒体汇刊，27: 2304–2319。Huang, T. 2024. 通过协调的迁移学习和模态对齐实现高效遥感。*CoRR*，abs/2404.18253。Kim, Y. 2014. 用于句子分类的卷积神经网络。载于 Moschitti, A.; Pang, B.; and Daelemans, W., 编，*2014*年自然语言处理实证方法会议论文集，*EMNLP 2014*，*2014*年*10*月*25*日至*29*日，多哈，卡塔尔，*ACL*特别兴趣小组*SIGDAT*会议，1746–1751。ACL。Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2：借助冻结图像编码器和大语言模型启动语言-图像预训练。载于国际机器学习会议，19730–19742。PMLR。Li, J.; Pan, J.; Sun, Y.; and Huang, X. 2025. 融合视觉-语言的语义感知船舶检测。arXiv:2508.15930。Li, J.; Wang, N.; Zhang, L.; Du, B.; and Tao, D. 2020a. 用于图像修复的循环特征推理。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，7760–7768。Li, Y.; Ma, J.; and Zhang, Y. 2021. 遥感大数据图像检索：综述。信息融合，67: 94–115。Li, Y.; Song, L.; Chen, Y.; Li, Z.; Zhang, X.; Wang, X.; and Sun, J. 2020b. 学习用于语义分割的动态路由。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，8553–8562。Lin, J.; Rao, Y.; Lu, J.; and Zhou, J. 2017. 运行时神经网络剪枝。神经信息处理系统进展，30。Lin, Z.; Wang, Z.; Qian, T.; Mu, P.; Chan, S.; and Bai, C. 2025. NeighborRetr：平衡跨模态检索中的枢纽中心性。载于*IEEE/CVF*计算机视觉与模式识别会议，*CVPR 2025*，美国田纳西州纳什维尔，*2025*年*6*月*11*日至*15*日，9263–9273。计算机视觉基金会/IEEE。

Liu, Y.; Pan, J.; Yang, J.; Chen, T.; Zhou, P.; and Zhang, B. 2025. 利用扩散模型生成多样实例，以增强遥感图像中的少样本目标检测。*IEEE*地球科学与遥感快报。Liu, Y.; Pan, J.; and Zhang, B. 2025. 可控复制粘贴：用于遥感少样本目标检测的可控扩散增强方法。arXiv:2507.21816。Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer：采用移位窗口的层次化视觉Transformer。载于*2021*年*IEEE/CVF*国际计算机视觉会议，*ICCV 2021*，加拿大魁北克省蒙特利尔，*2021*年*10*月*10*日至*17*日，9992–10002。IEEE。Lu, X.; Wang, B.; Zheng, X.; and Li, X. 2018. 探索遥感图像图注生成的模型与数据。*IEEE*地球科学与遥感汇刊，56(4): 2183–2195。Ma, Q.; Pan, J.; and Bai, C. 2024. 面向遥感图像-文本检索的方向导向视觉语义嵌入模型。*IEEE*地球科学与遥感汇刊。Ouyang, P.; Chen, J.; Ma, Q.; Wang, Z.; and Bai, C. 2024. 区分视觉相似图像：用于图像-文本检索的三元组对比学习框架。载于*IEEE*国际多媒体与博览会会议，*ICME 2024*，加拿大安大略省尼亚加拉瀑布城，*2024*年*7*月*15*日至*19*日，1–6。IEEE。Ouyang, P.; Ma, Q.; and Bai, C. 2025. 面向遥感跨模态检索的稀疏信息感知网络。*IEEE*地球科学与遥感汇刊，63: 1–15。Pan, J.; Lei, S.; Fu, Y.; Li, J.; Liu, Y.; Sun, Y.; He, X.; Peng, L.; Huang, X.; and Zhao, B. 2025a. EarthSynth：利用扩散模型生成信息丰富的地球观测数据。arXiv:2505.12108。Pan, J.; Liu, Y.; Fu, Y.; Ma, M.; Li, J.; Paudel, D. P.; Van Gool, L.; and Huang, X. 2025b. 在地球上定位万物：推动面向遥感领域的开放词汇目标检测。载于人工智能促进会会议论文集，第39卷，6281–6289。Pan, J.; Ma, Q.; and Bai, C. 2023a. 遥感图像-文本检索的先验指令表征框架。载于 El-Saddik, A.; Mei, T.; Cucchiara, R.; Bertini, M.; Vallejo, D. P. T.; Atrey, P. K.; and Hossain, M. S., 编，第*31*届*ACM*国际多媒体会议论文集，*MM 2023*，加拿大安大略省渥太华，*2023*年*10*月*29*日至*11*月*3*日，611–620。ACM。Pan, J.; Ma, Q.; and Bai, C. 2023b. 减少语义混淆：用于遥感跨模态检索的场景感知聚合网络。载于*2023*年*ACM*国际多媒体检索会议论文集，398–406。Qin, Y.; Peng, D.; Peng, X.; Wang, X.; and Hu, P. 2022. 基于噪声对应关系的跨模态检索深度证据学习。载于*ACM*多媒体。Qu, B.; Li, X.; Tao, D.; and Lu, X. 2016. 高分辨率遥感图像的深层语义理解。

载于计算机、信息与电信系统国际会议，*CITS 2016*，中国昆明，*2016*年*7*月*6*日至*8*日，1–5。IEEE。Qu, L.; Liu, M.; Wu, J.; Gao, Z.; and Nie, L. 2021. 面向图文检索的动态模态交互建模。载于第*44*届国际*ACM SIGIR*信息检索研究与发展会议论文集，1104–1113。Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. FaceNet：用于人脸识别和聚类的统一嵌入。载于*IEEE*计算机视觉与模式识别会议，*CVPR 2015*，美国马萨诸塞州波士顿，*2015*年*6*月*7*日至*12*日，815–823。IEEE计算机学会。Tang, X.; Wang, Y.; Ma, J.; Zhang, X.; Liu, F.; and Jiao, L. 2023. 用于跨模态遥感图像与文本检索的交互增强特征变换器。*IEEE*地球科学与遥感汇刊，61：1–15。Wang, F.; Pan, J.; Xu, S.; and Tang, J. 2022. 面向RGB-D显著性检测的判别性跨模态特征学习。*IEEE*图像处理汇刊，31：1285–1297。Wright, J.; Ma, Y.; Mairal, J.; Sapiro, G.; Huang, T. S.; and Yan, S. 2010. 面向计算机视觉与模式识别的稀疏表示。*IEEE*会刊，98(6)：1031–1044。Wu, Y.; Li, L.; Jiao, L.; Liu, F.; Liu, X.; and Yang, S. 2024. TrTr-CMR：用于遥感图像描述的跨模态推理双变换器。*IEEE*地球科学与遥感汇刊。Yang, S.; Xu, Z.; Wang, K.; You, Y.; Yao, H.; Liu, T.; and Xu, M. 2023. BiCro：通过双向跨模态相似性一致性纠正多模态数据中的噪声对应关系。载于*IEEE/CVF*计算机视觉与模式识别会议，*CVPR 2023*，加拿大不列颠哥伦比亚省温哥华，*2023*年*6*月*17*日至*24*日，19883–19892。IEEE。Yu, D.; and Ji, S. 2022. 一种面向遥感图像的新型空间导向目标检测框架。*IEEE*地球科学与遥感汇刊，60：1–16。Yuan, Z.; Zhang, W.; Fu, K.; Li, X.; Deng, C.; Wang, H.; and Sun, X. 2022. 探索一种用于跨模态遥感图像检索的细粒度多尺度方法。*IEEE*地球科学与遥感汇刊，60：1–19。Zhang, D.; Nan, F.; Wei, X.; Li, S.; Zhu, H.; McKeown, K.; Nallapati, R.; Arnold, A.; and Xiang, B. 2021. 通过对比学习支持聚类。*arXiv*预印本 *arXiv:2103.12953*。Zhang, L.; Xu, D.; Arnab, A.; and Torr, P. H. 2020. 动态图消息传递网络。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，3726–3735。Zhang, L.; and Zhang, L. 2022. 面向遥感数据分析的人工智能：挑战与机遇综述。*IEEE*地球科学与遥感杂志，10(2)：270–294。Zhang, Z.; Zhao, T.; Guo, Y.; and Yin, J. 2024. RS5M和GeoRSCLIP：用于遥感的大规模视觉语言数据集与大型视觉语言模型。*IEEE*地球科学与遥感汇刊。

Zhong, G.; Huan, R.; Wu, M.; Liang, R.; and Chen, P. 2025. 面向模态缺失与分布偏移下的鲁棒多模态情感识别。*arXiv*预印本 *arXiv:2506.10452*。Zhou, J.; Jampani, V.; Pi, Z.; Liu, Q.; and Yang, M.-H. 2021. 解耦动态滤波器网络。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，6647–6656。Zhou, Y.; Suo, J.; Wang, Y.; Su, J.; Xiao, W.; Hong, Z.; Ranjan, R.; Wang, L.; and Wen, Z. 2025. MMCANet：一种用于云去除并探索渐进式遥感图像复原算法的多模态交叉注意力网络。*IEEE*地球科学与遥感汇刊。