南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

DiffCLIP: Few-shot Language-driven Multimodal Classifier

Jiaqing Zhang[1*], Mingxiang Cao[1*], Xue Yang[2], Kai Jiang[1], Yunsong Li[1†]

[1]The State Key Laboratory of Integrated Services Networks, Xidian University
[2]Shanghai AI Laboratory

汇报人：蒋明忠

时间：2025.09

AAAI

Background

最近,对比语言图像预训练框架CLIP取得了显著的成功,为语义分割、目标检测和3D点云理解等后续任务提供了基础。

然而,这些模型在应用于特定领域(如遥感图像)时往往会遇到困难。这是因为这些模型是在自然图像上训练的, 可能无法完全捕捉到特定领域的多样性和复杂性。为了解决这个问题,大多数研究都集中在为每个领域构建大规模的预训练数据集,并进行额外的微调阶段,以适应医疗、电子商务和遥感等领域的下游任务。

大规模高维多模态遥感图像的专业可用性限制了高维图像的监督学习。因此,一个自然的问题出现了:我们能否避免收集和标记数据的成本,将CLIP引入到标记样本较少的高维多模态图像分类中? 遗憾的是,利用无监督学习知识来处理CLIP的Few-shot训练还没有得到充分的探索。

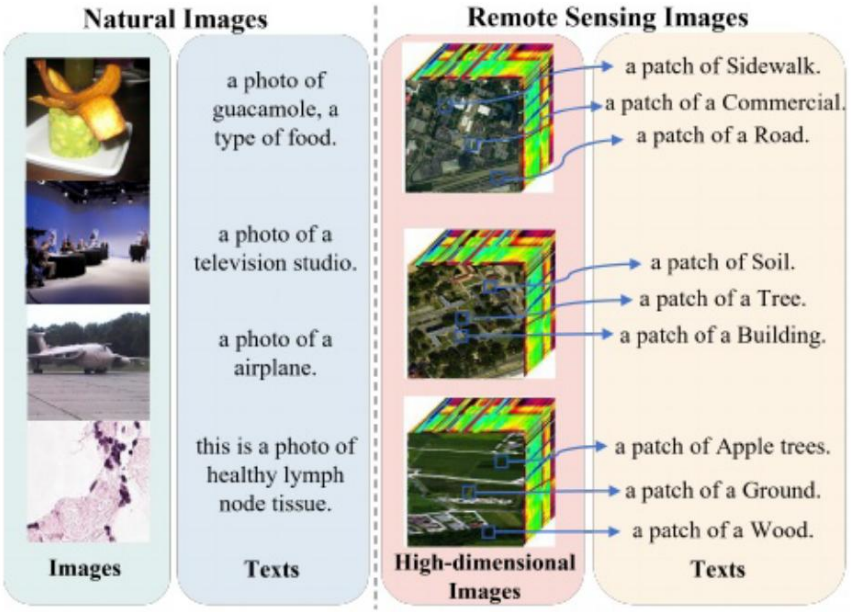
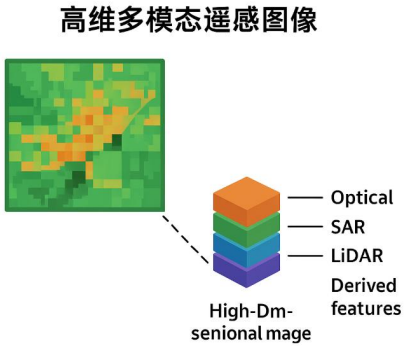


图1:(a) CLIP模型使用自然图像数据集中的四个随机采样的图像-文本对进行训练，这些数据集中有丰富的标记样例。(b)遥感图像以补丁格式注释，但由于遥感注释的特殊性，注释的补丁样本严重缺乏。需要专业的专业知识和高效的时间管理。数据集之间的这种显著差异使得将CLIP模型直接应用于遥感应用具有挑战性。



Background

同一地区的多模态遥感数据（如高光谱、LiDAR等）能够提供互补的地面特征，通过联合分类可显著提高精度，因此被广泛用于城市规划、自然资源管理和环境监测等领域。

高光谱成像(HSI)因其高维光谱信息可实现基于反射率的材料识别，为遥感分类提供了多维模态支持。

关键在于如何有效整合和学习不同模态的特征，以捕捉跨模态的共享信息并减少模态差距。已有研究通过共享子空间或共享流形模型实现特征对齐，近期的Transformer和扩散模型则将多模态分类扩展到更大规模。然而，这些模型大多局限于视觉维度的一致性，缺乏视觉与语言联合探索的视角。

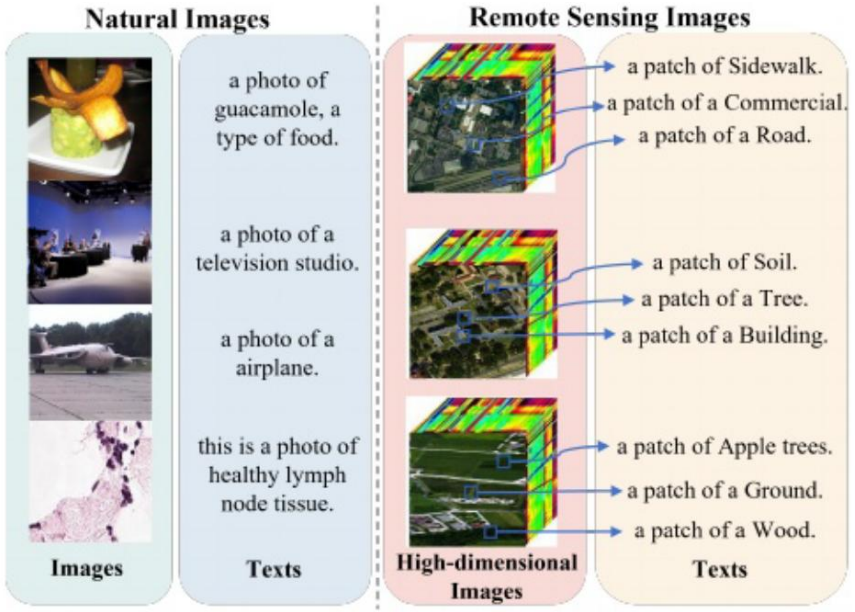
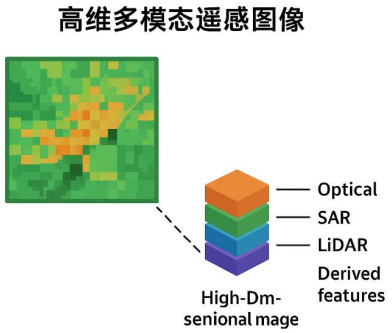


图1:(a) CLIP模型使用自然图像数据集中的四个随机采样的图像-文本对进行训练，这些数据集中有丰富的标记样例。(b)遥感图像以补丁格式注释，但由于遥感注释的特殊性，注释的补丁样本严重缺乏。需要专业的专业知识和高效的时间管理。数据集之间的这种显著差异使得将CLIP模型直接应用于遥感应用具有挑战性。



Background

对比语言-图像预训练(CLIP)等视觉语言模型在分析带有语言信息的自然图像方面表现出色。

然而,由于用于训练的图像文本对的可用性有限,这些模型在应用于诸如遥感等专门领域时经常遇到挑战。

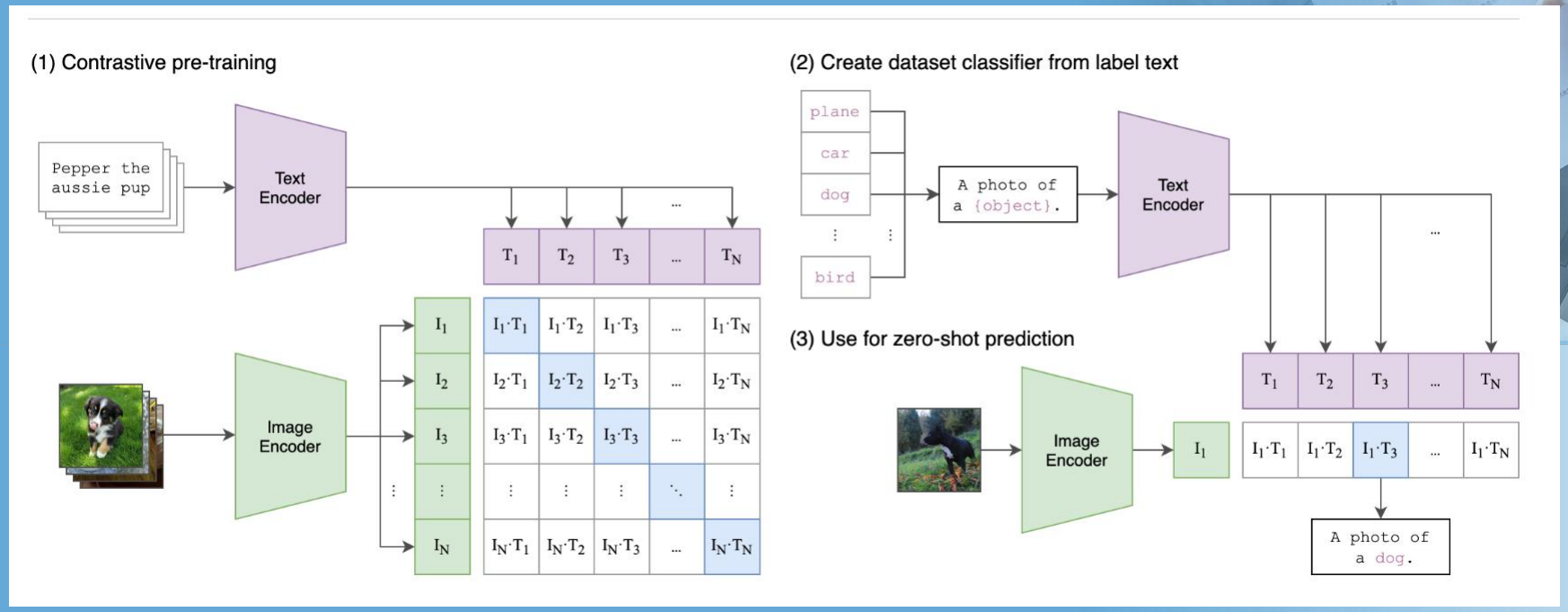
为了解决这个问题,我们引入了一个新的框架DiffCLIP,它扩展了CLIP,以有效地传递全面的语言驱动语义信息,从而实现高维多模态遥感图像的准确分类。

DiffCLIP是一种少量学习方法,它利用未标记的图像进行预训练。它采用无监督掩模扩散学习,在不需要标记的情况下捕获不同模态的分布。模态共享的图像编码器将多模态数据映射到统一的子空间中,在模态之间提取具有一致参数的共享特征。训练有素的图像编码器通过将视觉表示与来自CLIP的类标签文本信息对齐来进一步增强学习。

Method

CLIP 三步走：

1. 图文对比学习，让图和文在一个空间里。
2. 用标签生成文本提示，直接当分类器。
3. 新图进来，跟文本比相似度，谁最近就判谁。



Method

该方法包括两个关键阶段:无监督学习和few-shot学习。 无监督学习阶段涉及正向掩模扩散和反向去噪恢复。为 了降低训练成本,只有数据集的一个子集进行前向掩模扩散,而反向过程使用共享图像编码器来学习多模态特征,由两个模态特定的解码器支持。 Few-shot学习阶段采用语言驱动的监督,提供比离散标签更丰富的语义上下文, 能够更好地捕获复杂数据的细微差别并改进分类。

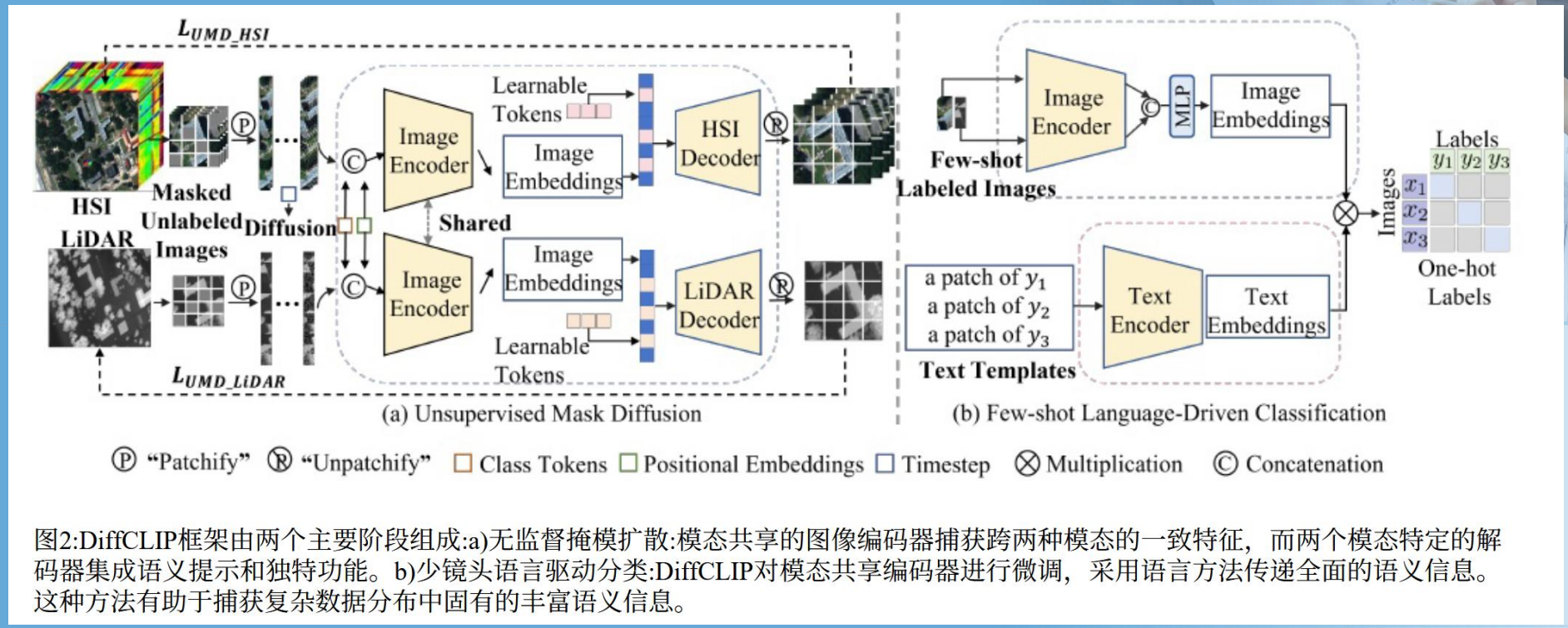


图2:DiffCLIP框架由两个主要阶段组成:a)无监督掩模扩散:模态共享的图像编码器捕获跨两种模态的一致特征，而两个模态特定的解码器集成语义提示和独特功能。b)少镜头语言驱动分类:DiffCLIP对模态共享编码器进行微调，采用语言方法传递全面的语义信息。这种方法有助于捕获复杂数据分布中固有的丰富语义信息。

Method

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

编码器-无监督掩模扩散:

在 **DiffCLIP** 中，正向掩模扩散从干净样本

$$x_0 \sim Q(x_0)$$

开始。对每个模态，采用 **不对称掩模策略**，将输入划分为

- **掩蔽区域** x_0^m
- **可见区域** x_0^v

然后，**只对可见区域** x_0^v **进行扩散**：在每一步 $t = 1, \ldots, T$，对可见区域递归加入少量高斯噪声，得到

$$x_1^v, x_2^v, \ldots, x_T^v,$$

其演化遵循马尔可夫过程：

$$\mathcal{Q}(x_t^v | x_{t-1}^v) = \mathcal{N}(x_t^v; \sqrt{1-\beta_t} x_{t-1}^v, \beta_t I), \quad (1)$$

$$\mathcal{Q}(x_1^v, \ldots, x_T^v | x_0^v) = \prod_{t=1}^{T} \mathcal{Q}(x_t^v | x_{t-1}^v), \quad (2)$$

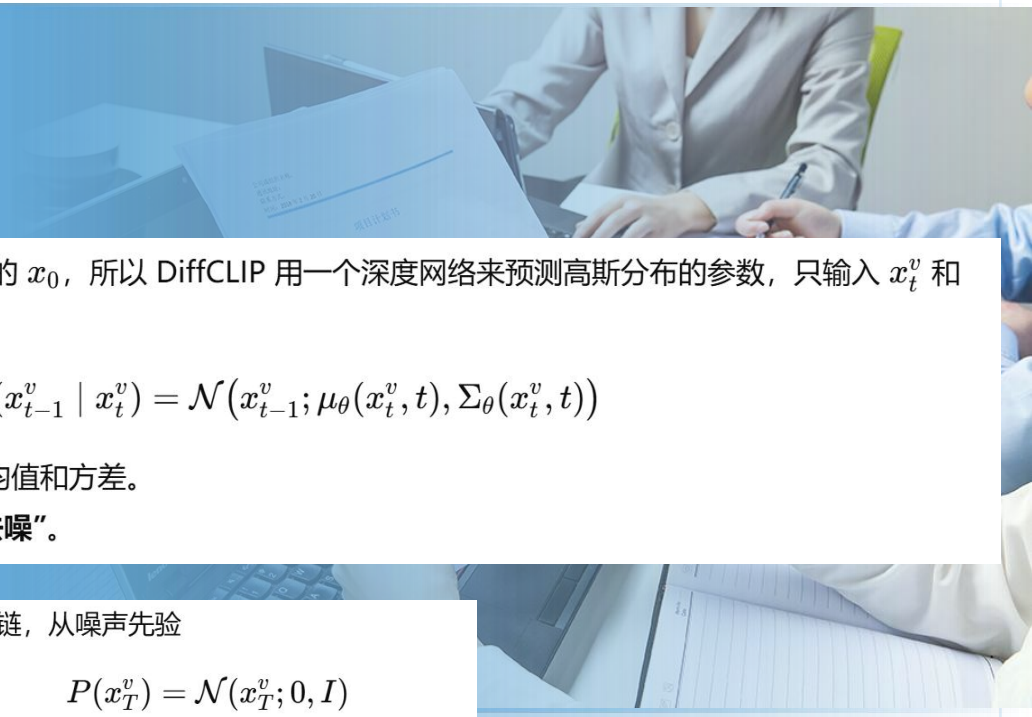
$$\hat{x}^v = \mathcal{E}_\theta \left(\{ p_{ct} + p_p + p_{dt}; x^v \}_1 \right), \quad (7)$$

直观上，掩蔽区域 x_0^m 保持不变，可见区域 x_0^v 被逐步加噪。这样模型在训练中既能利用未掩蔽的局部信息，也被迫学习掩蔽区域与可见区域的**跨模态共享特征**，从而增强多模态表征能力。

Method

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

解码器-反向去噪恢复:



1. 反向过程的基本思想

DiffCLIP 在正向过程中对可见区域 x_0^v 逐步加噪得到 x_T^v。
反向过程要做的就是**从噪声状态 x_T^v 逐步预测回干净状态 x_0^v**。

2. 公式(3)：贝叶斯后验

当噪声方差 β_t 足够小时，正向扩散的后验分布

$$Q(x_{t-1} \mid x_t, x_0)$$

可以看成高斯分布：

$$Q(x_{t-1} \mid x_t, x_0) = \mathcal{N}\big(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I\big)$$

也就是说，如果知道 x_t 和原始样本 x_0，就可以直接算出 x_{t-1} 的高斯分布。

实际反向采样时我们不知道真实的 x_0，所以 DiffCLIP 用一个深度网络来预测高斯分布的参数，只输入 x_t^v 和时间 t：

$$P_\theta(x_{t-1}^v \mid x_t^v) = \mathcal{N}\big(x_{t-1}^v; \mu_\theta(x_t^v, t), \Sigma_\theta(x_t^v, t)\big)$$

其中 μ_θ 和 Σ_θ 就是网络输出的均值和方差。
这相当于**网络学习“每一步怎么去噪”**。

整个反向过程就是一个高斯马尔可夫链，从噪声先验

$$P(x_T^v) = \mathcal{N}(x_T^v; 0, I)$$

开始，按照网络预测的条件分布一步步反向采样：

$$P_\theta(x_{0:T}^v) = P(x_T^v) \prod_{t=1}^{T} P_\theta(x_{t-1}^v \mid x_t^v)$$

这就是反向扩散的联合分布。

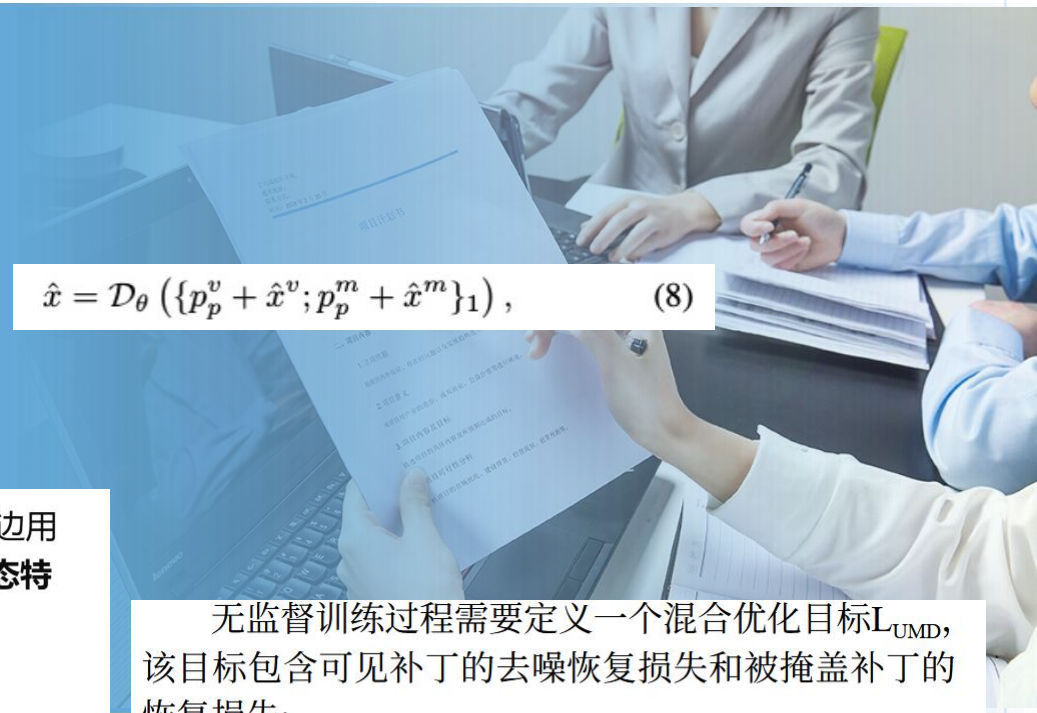
Method

解码器-反向去噪恢复:

DiffCLIP 在反向扩散过程中输出去噪预测 $f(x_T, t)$。
训练目标是最小化它与真实干净样本 x_0 的均方误差：

$$\mathbb{E}_{x_0 \sim Q(x_0)} \, \mathbb{E}_{T,\epsilon} \left\| x_0 - f(x_T, t) \right\|^2$$

这就是**去噪恢复损失**，用来更新整个模型。



$$\hat{x} = \mathcal{D}_\theta \left(\{ p_p^v + \hat{x}^v ; p_p^m + \hat{x}^m \}_1 \right), \quad (8)$$

> **DiffCLIP** 先把原始数据划分为可见部分和被掩蔽部分，然后一边训练网络去"迭代去噪"可见部分、一边用一组可学习的掩码令牌恢复被掩蔽部分。最后用**一个共享的图像编码器**提取跨模态语义，再用**两个模态特定的解码器**重建各自模态的细节，通过最小化去噪重建损失来更新整个模型。

换句话说，**三步走**：

1. 对未掩蔽 patch 进行迭代去噪、增强预测能力；
2. 用可学习的掩码向量重建被掩蔽 patch；
3. 用共享编码器 + 模态特定解码器，最小化去噪重建损失来训练 DiffCLIP。

无监督训练过程需要定义一个混合优化目标L_{UMD}，该目标包含可见补丁的去噪恢复损失和被掩盖补丁的恢复损失:

$$\mathcal{L}_{\text{UMD}} = \mathbb{E}_{x_0} \| x_0 - \hat{x} \|^2. \quad (9)$$

Method

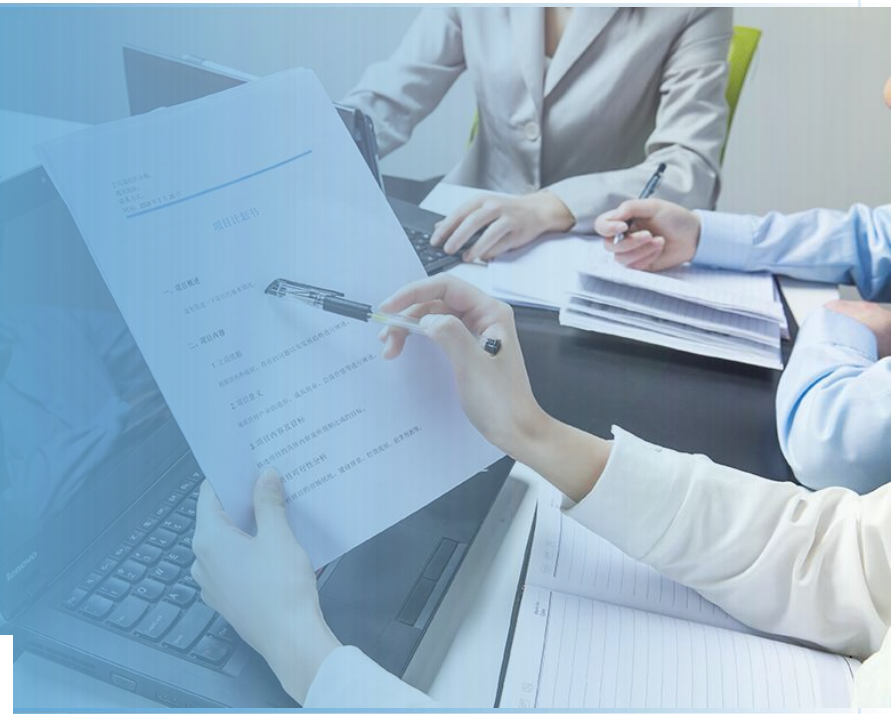
Few-shot Language-Driven Classification:

DiffCLIP的训练目标涉及两个分类任务:

给定图像p(y | x)预测文本和给定文本p(x |y)预测图像。批处理中的每个样本被分配一个与其配对数据相对应的标签作为目标。

在这一部分，作者通过生成针对特定类别的文本描述来增强模型的语义理解，而不是简单加一些形容词，也没有使用任何预训练的文本编码器权重，从而验证方法本身的有效性。

具体做法是：利用 GPT-4 的先验知识，为每个类别生成包含固有属性、类间关系和类名的描述。



p1 基线
"草"

p2 长 + 弱相关
"休斯顿城市阳光下的一片草地公园与花园景象"

p3 短融合
"草地地类"

p4 长 + 强相关
"光合作用旺盛的绿色草冠层，NDVI大于0.5，高度低于20厘米，定期灌溉，近红外波段高反射"

p5 类特定描述
"草坪草，NDVI 0.52，株高18厘米，已灌溉，纹理平滑，近红外反射高，激光雷达高程低"

Method

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Few-shot Language-Driven Classification:

在编码之前,我们通过简单的嵌入获得文本信息的标记化表示。随后, 我们利用转换器对标记化表示进行编码并产生文本特 征嵌入,并且将嵌入归一化为单位范数。

$$v_{\text{text}} = \mathcal{T}_\theta(y), z_{\text{text}} = v_{\text{text}} / \|v_{\text{text}}\|, \quad (10)$$

我们假设模态共享编码器已经有效地捕获了两个模态之间的公共信息。 此外,在解码过程中,它通过去噪恢复任务被注入了鲁棒的语义线索,同时保留了每个模态的独特特征。

$$v_{\text{image}} = \mathcal{E}_\theta(x), z_{\text{image}} = v_{\text{image}} / \|v_{\text{image}}\|, \quad (11)$$

$$p(x \mid y) = \sigma_\tau \left(z_{\text{text}}, \left\{ z^i_{\text{image}} \right\}_{i=1}^N \right) \in \mathbb{R}^N, \quad (12)$$

$$p(y \mid x) = \sigma_\tau \left(z_{\text{image}}, \left\{ z^i_{\text{text}} \right\}_{i=1}^N \right) \in \mathbb{R}^N, \quad (13)$$

$$p\left(x = x_i \mid y\right) = \frac{\exp\left(z_{\text{text}} \cdot z^i_{\text{image}} / \tau\right)}{\sum_{j=1}^N \exp\left(z_{\text{text}} \cdot z^j_{\text{image}} / \tau\right)}, \quad (14)$$

$$p\left(y = y_i \mid x\right) = \frac{\exp\left(z_{\text{image}} \cdot z^i_{\text{text}} / \tau\right)}{\sum_{j=1}^N \exp\left(z_{\text{image}} \cdot z^j_{\text{text}} / \tau\right)}. \quad (15)$$

$$\mathcal{L}_{\text{FLC}} = \frac{1}{2N} \sum_{i=1}^{N} \left(H\left(p\left(y \mid x_i\right), \mathbf{e}_i\right) + H\left(p\left(x \mid y_i\right), \mathbf{e}_i\right) \right), \quad (16)$$

Method

该方法包括两个关键阶段:无监督学习和few-shot学习。 无监督学习阶段涉及正向掩模扩散和反向去噪恢复。为 了降低训练成本,只有数据集的一个子集进行前向掩模扩散,而反向过程使用共享图像编码器来学习多模态特征,由两个模态特定的解码器支持。 Few-shot学习阶段采用语言驱动的监督,提供比离散标签更丰富的语义上下文, 能够更好地捕获复杂数据的细微差别并改进分类。

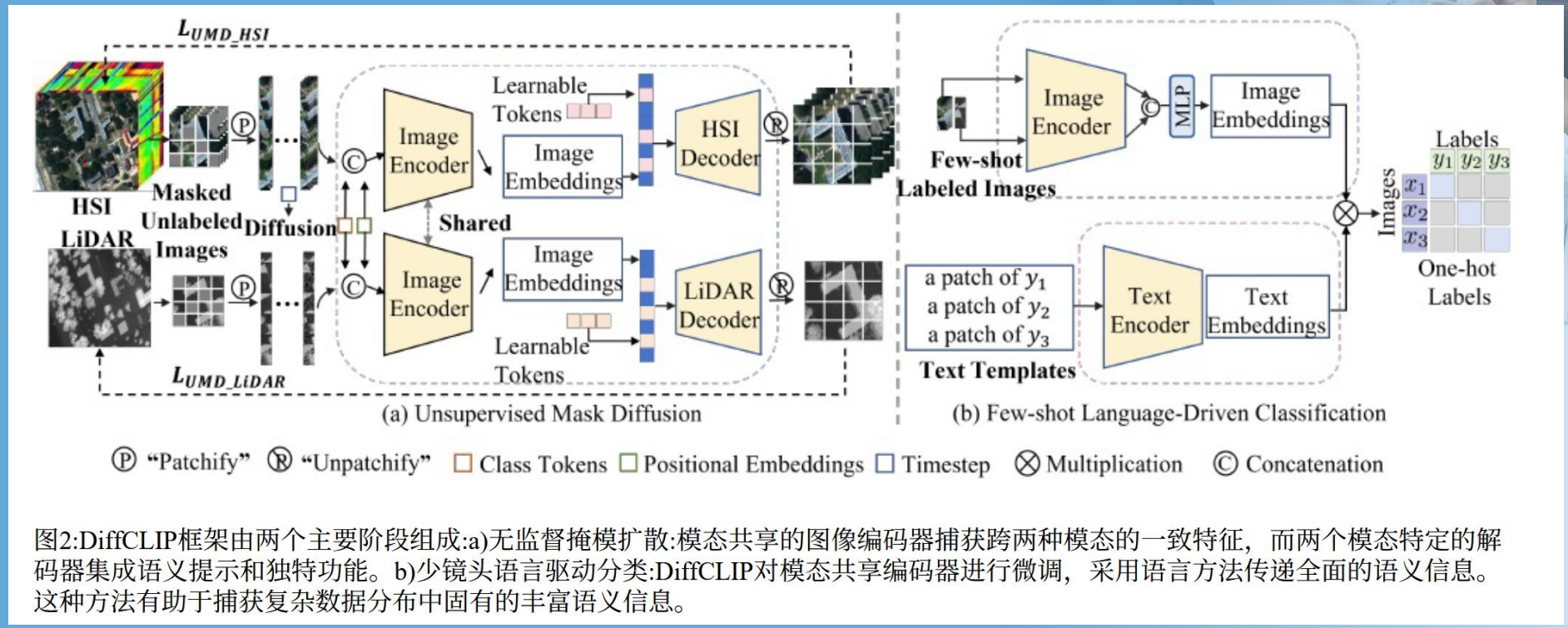


图2:DiffCLIP框架由两个主要阶段组成:a)无监督掩模扩散:模态共享的图像编码器捕获跨两种模态的一致特征，而两个模态特定的解码器集成语义提示和独特功能。b)少镜头语言驱动分类:DiffCLIP对模态共享编码器进行微调，采用语言方法传递全面的语义信息。这种方法有助于捕获复杂数据分布中固有的丰富语义信息。

Experiments

作者在四个广泛认可的多模态遥感基准数据集（Houston、Trento、MUUFL 和 MRNet）上评估了所提出方法的性能，并采用总体准确率（OA）、平均准确率（AA）和 Kappa 系数三种指标进行定量评价。为验证方法有效性，选取了四种多模态遥感分类的 SOTA 方法（GLT、CALC、MIViT、LDS2AE）、五种少镜头学习的 SOTA 算法（RN-FSC、MFRNML、SC-Former、HIPL、MMPR）以及 CLIP 作为对比。实验在 NVIDIA GeForce RTX A100 GPU 上进行，训练样本预处理为 11×11 补丁，优化采用 Adam（初始学习率 1e-4、权重衰减 1e-5），并使用余弦调度器（无监督学习）与阶跃调度器（少镜头学习）。

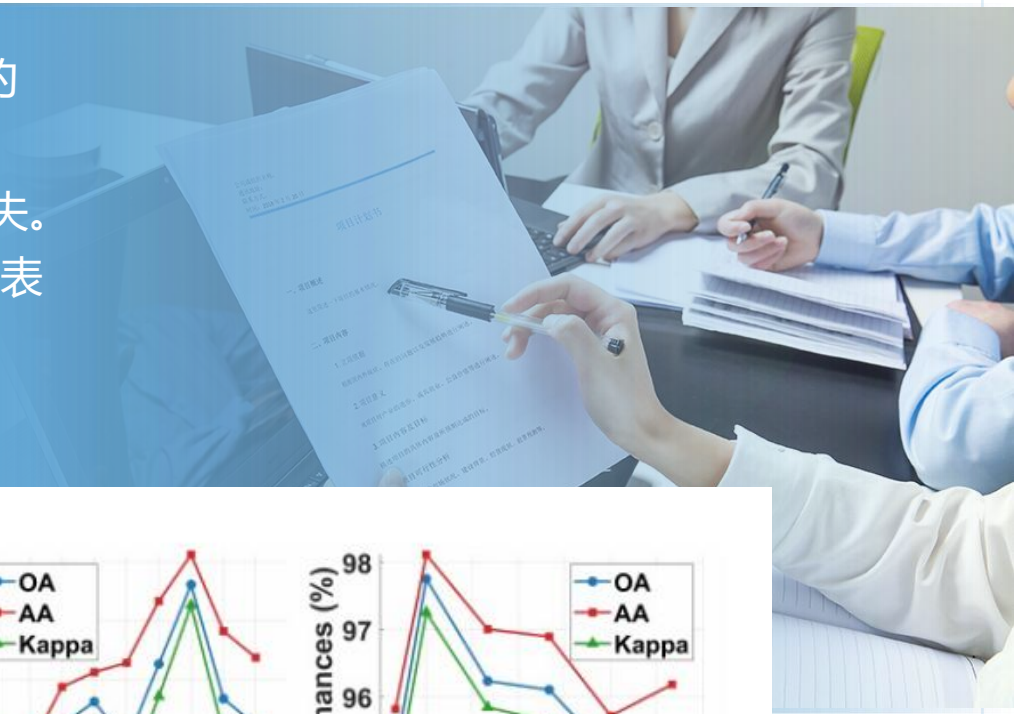
Settings	Methods	Houston			MUUFL			Trento		
		OA(%)	AA(%)	Kappa(%)	OA(%)	AA(%)	Kappa(%)	OA(%)	AA(%)	Kappa(%)
2-shot	CLIP	35.83	42.75	32.44	51.62	38.26	38.91	85.33	78.63	82.47
	LDS^{2}AE	48.68	51.68	44.85	56.47	44.86	43.78	90.23	79.25	86.76
	MFRN-ML	48.73	51.55	43.79	57.16	44.54	43.12	90.81	79.33	86.04
	SCFormer	49.09	52.93	46.11	57.37	46.94	45.13	91.55	80.05	88.24
	HIPL	50.96	54.33	47.15	58.84	52.16	47.37	92.11	80.14	88.59
	MMPR	49.79	53.01	46.78	57.86	49.05	46.15	91.75	79.88	87.93
	DiffCLIP	**52.15**	**56.02**	**48.39**	**59.39**	**54.94**	**48.80**	**93.19**	**80.45**	**90.85**
8-shot	CLIP	52.11	56.37	52.86	67.56	66.78	63.21	92.76	91.66	92.53
	LDS^{2}AE	57.72	57.60	54.36	70.59	68.43	62.94	95.49	95.62	94.03
	MFRN-ML	58.03	60.59	55.71	72.36	70.28	65.39	95.54	95.78	94.02
	SCFormer	59.61	60.92	56.25	74.38	70.64	67.82	95.57	95.79	94.18
	HIPL	60.36	63.89	57.54	75.65	72.63	69.15	95.98	95.91	94.79
	MMPR	59.98	62.57	56.92	74.67	71.68	67.98	95.63	95.72	94.33
	DiffCLIP	**61.93**	**65.93**	**58.95**	**77.60**	**74.97**	**71.53**	**96.30**	**96.22**	**95.09**
20-shot	CLIP	56.31	59.98	54.32	67.30	71.79	65.10	91.28	93.11	92.43
	LDS^{2}AE	65.71	70.50	63.09	78.13	77.24	72.01	98.13	96.91	97.50
	MFRN-ML	74.49	77.81	73.22	78.83	77.45	72.32	98.25	96.98	97.43
	SCFormer	74.97	78.16	73.80	79.11	78.55	73.23	98.27	96.95	97.36
	HIPL	79.36	82.11	76.35	80.87	79.36	76.28	98.41	97.15	97.86
	MMPR	77.93	80.26	75.70	80.02	78.71	74.63	98.40	97.03	97.42
	DiffCLIP	**81.87**	**84.09**	**80.40**	**81.81**	**80.67**	**76.64**	**98.60**	**97.90**	**98.13**

表1:不同few-shot设置下三个数据集的对比结果。

Experiments

通过改变样本数量从300到700来评估无监督掩膜扩散阶段样本大小的影响。结果表明,增加样本量可显著改善OA、AA和Kappa系数。

增加掩蔽比最初通过减少冗余来提高性能,但过度掩蔽会导致信息丢失。结果显示在70% 处出现峰值,这是为进一步的实验选择的。同样,消融 实验表明,贴片大小为11表现最佳,平衡类别分离和接受野大小以获得最佳结果。



Sample numbers	OA(%)	AA(%)	Kappa(%)
300	91.75	91.11	91.62
500	92.58	91.76	93.00
700	**94.80**	**94.24**	**93.28**

Settings	OA(%)	AA(%)	Kappa(%)
w/o Text	89.35	89.66	89.01
w/o Diffusion	89.67	90.73	89.28
w/o Mask	90.78	91.72	89.83
w/o Unsupervised	87.82	88.94	87.76
DiffCLIP	**94.80**	**94.24**	**93.28**

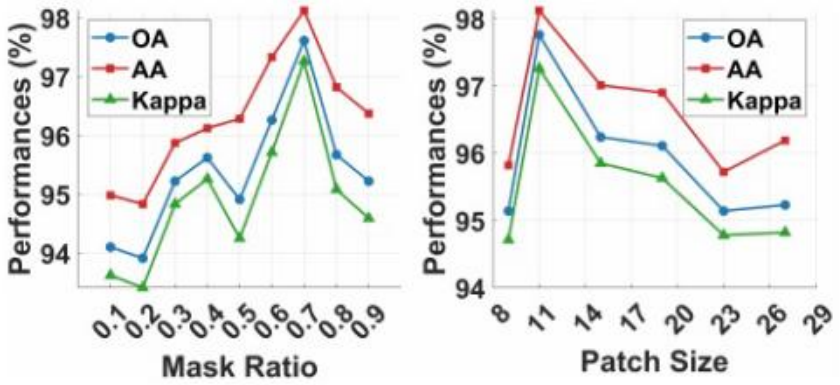


图4:不同掩蔽比和补丁大小的休斯顿数据集分类性能。

Experiments

作者在Trento数据集上通过图 3 的分类结果进行了定性评估，结果显示 DiffCLIP 在视觉分类中表现最佳，充分利用语义信息，在道路等类别上呈现出连续、准确的分类效果。

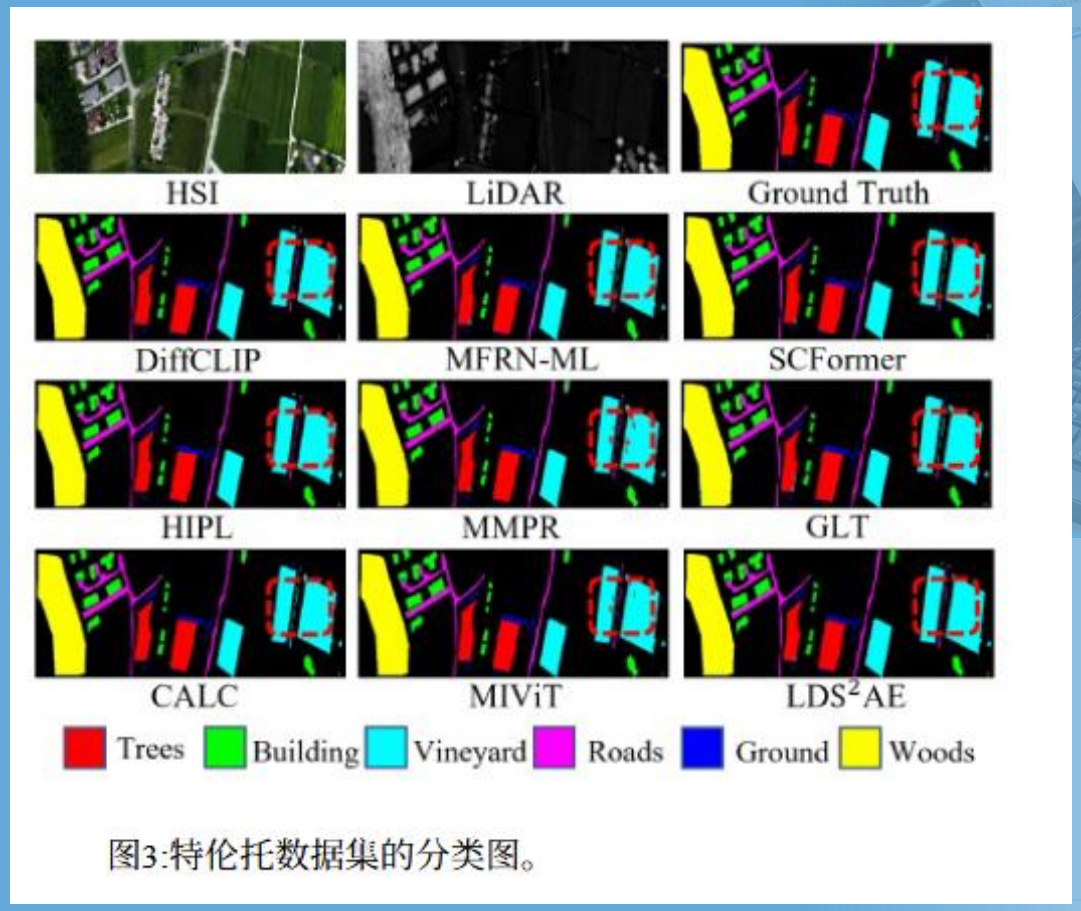


图3:特伦托数据集的分类图。

Experiments

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

作者通过在休斯顿数据集上进行消融实验，评估文本提示（prompts）对 DiffCLIP 分类性能的影响。他们设计了五组不同的提示，与基线 p1 比较后发现：p2 的长文本因与任务相关性弱导致准确率下降；p3 对提示进行微调后有轻微提升；p4 表明更长且相关性强的提示能增强监督；p5 使用特定类别描述使 OA 准确率提高了 0.54%。这一结果表明，优化提示的长度与内容对提升 DiffCLIP 的性能非常关键。

	Context length	Text prompts	OA(%)	AA(%)	Kappa(%)
p1	5∼7	a patch of a {class name}.	97.61	98.12	97.26
p2	6∼8	a nice patch of a {class name}.	97.05	97.63	96.81
p3	7∼9	a fusion patch of a {class name}.	97.63	98.11	97.34
p4	11∼13	a multimodal fusion patch of a {class name} with strong semantic information.	97.93	98.51	97.64
p5	15∼30	specific class text descriptions.	**98.15**	**98.87**	**98.09**

p1 基线
"**草**"

p2 长 + 弱相关
"**休斯顿城市阳光下的一片草地公园与花园景象**"

p3 短融合
"**草地地类**"

p4 长 + 强相关
"**光合作用旺盛的绿色草冠层，NDVI大于0.5，高度低于20厘米，定期灌溉，近红外波段高反射**"

p5 类特定描述
"**草坪草，NDVI 0.52，株高18厘米，已灌溉，纹理平滑，近红外反射高，激光雷达高程低**"

Methods		Houston			MUUFL			Trento		
		OA(%)	AA(%)	Kappa(%)	OA(%)	AA(%)	Kappa(%)	OA(%)	AA(%)	Kappa(%)
Supervised	GLT	90.13	90.42	89.42	82.75	75.70	78.67	98.19	97.75	98.04
	CALC	87.87	88.92	86.87	81.94	64.09	77.01	97.11	92.31	96.64
	MIViT	93.21	93.87	92.75	83.13	79.74	78.91	98.03	97.96	98.24
	LDS^{2}AE	94.88	95.31	94.46	84.82	82.19	80.42	98.77	98.11	98.42
Few-shot	RN-FSC	93.42	94.53	93.64	84.83	74.52	81.25	98.69	97.66	98.61
	MFRN-ML	94.50	95.38	93.99	84.78	77.03	81.07	98.37	97.71	98.49
	SCFormer	95.96	96.25	94.84	84.93	79.49	81.39	98.52	97.86	98.55
	HIPL	96.83	97.01	96.75	85.45	80.71	82.18	98.81	98.15	98.65
	MMPR	96.54	96.77	95.98	85.07	80.36	81.57	98.43	97.79	98.46
	DiffCLIP	**98.15**	**98.87**	**98.09**	**86.98**	**85.01**	**82.81**	**99.26**	**98.84**	**98.95**

表5:OA、AA和Kappa在三个数据集上的比较结果。

Method	ACC	AUC	SE	SP
ELNet	0.639	0.703	0.624	0.650
TransMed-S	0.667	0.705	0.635	0.664
SSL-DcGaR	0.731	0.758	0.723	0.734
DiffCLIP	**0.763**	**0.787**	**0.750**	**0.756**

表6:MRNet数据集的比较结果。

Thanks