COIDO: Efficient Data Selection for Visual Instruction Tuning via Coupled Importance-Diversity Optimization

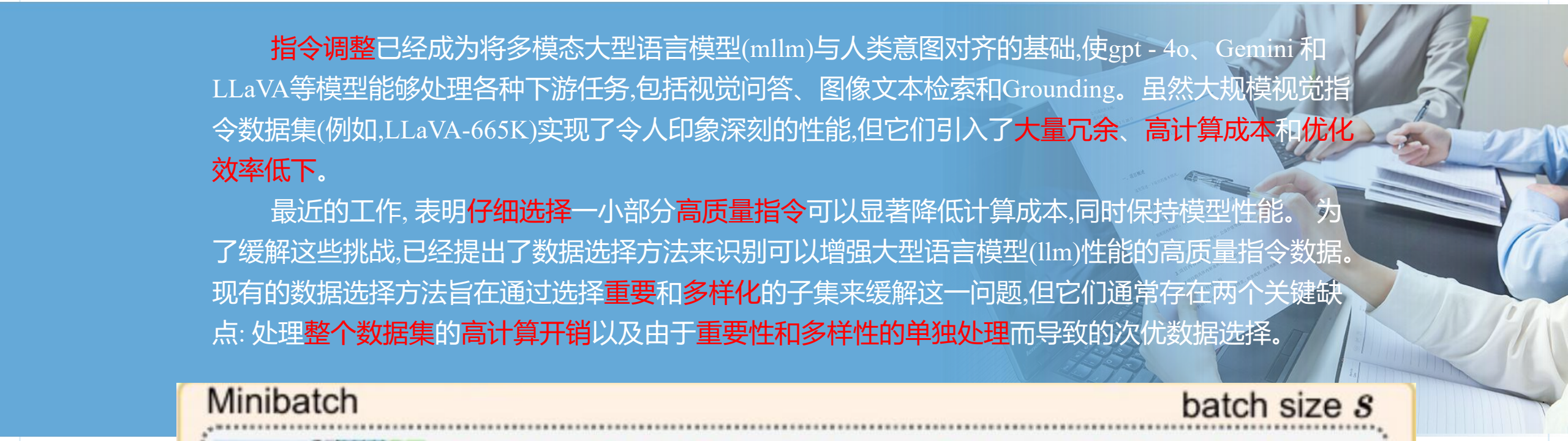
Yichen Yan[1,2], **Ming Zhong**[1,2], **Qi Zhu**[1], **Xiaoling Gu**[3], **Jinpeng Chen**[4], **Huan Li**[1,2*]
[1] The State Key Laboratory of Blockchain and Data Security, Zhejiang University
[2] Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security
[3] Hangzhou Dianzi University, Hangzhou, China
[4] School of Computer Science (National Pilot Software Engineering School), BUPT
{yichen.yan, chime, lihuan.cs}@zju.edu.cn,
qizhu.zju.research@gmail.com, guxl@hdu.edu.cn, jpchen@bupt.edu.cn

汇报人：蒋明忠

时间：2025.12

NeurIPS

Background

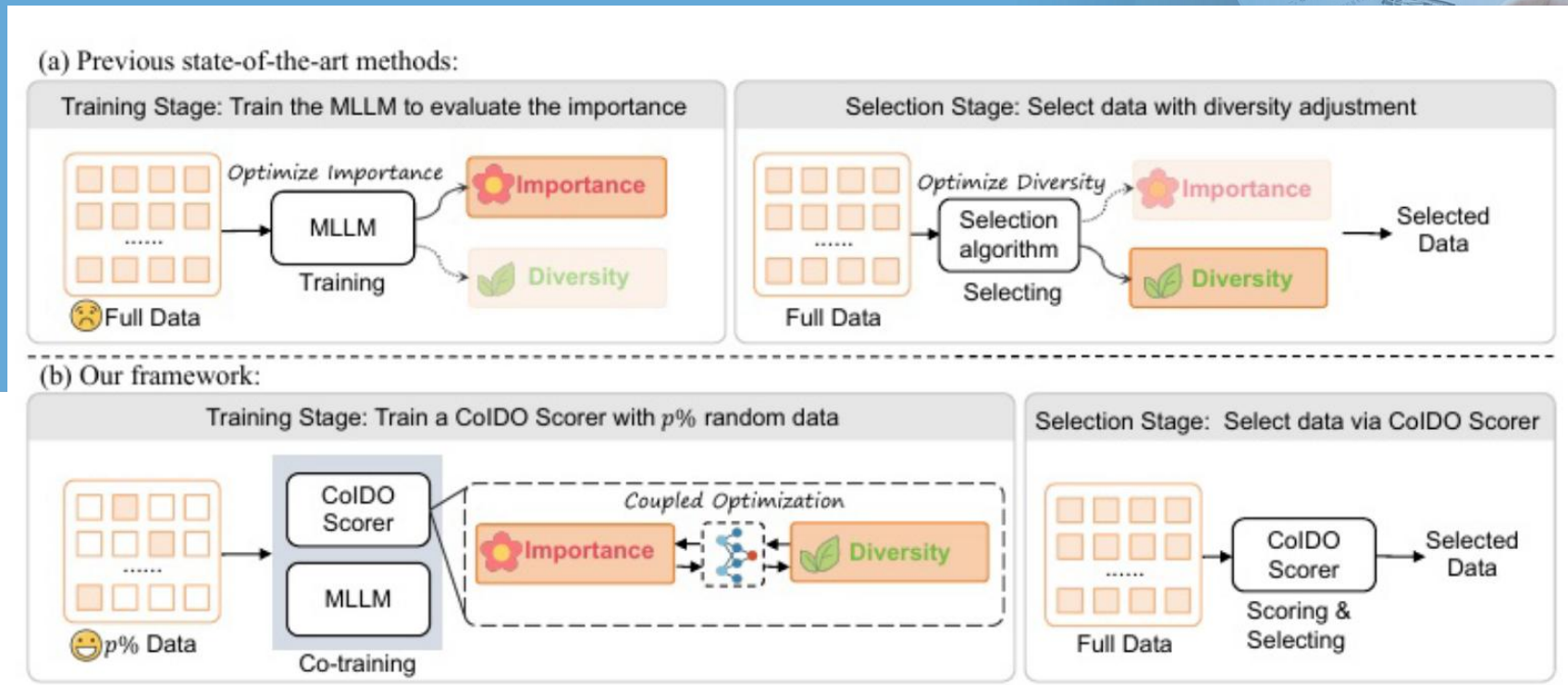


指令调整已经成为将多模态大型语言模型(mllm)与人类意图对齐的基础,使gpt - 4o、 Gemini 和LLaVA等模型能够处理各种下游任务,包括视觉问答、图像文本检索和Grounding。虽然大规模视觉指令数据集(例如,LLaVA-665K)实现了令人印象深刻的性能,但它们引入了大量冗余、高计算成本和优化效率低下。

最近的工作, 表明仔细选择一小部分高质量指令可以显著降低计算成本,同时保持模型性能。 为了缓解这些挑战,已经提出了数据选择方法来识别可以增强大型语言模型(llm)性能的高质量指令数据。现有的数据选择方法旨在通过选择重要和多样化的子集来缓解这一问题,但它们通常存在两个关键缺点: 处理整个数据集的高计算开销以及由于重要性和多样性的单独处理而导致的次优数据选择。

Background

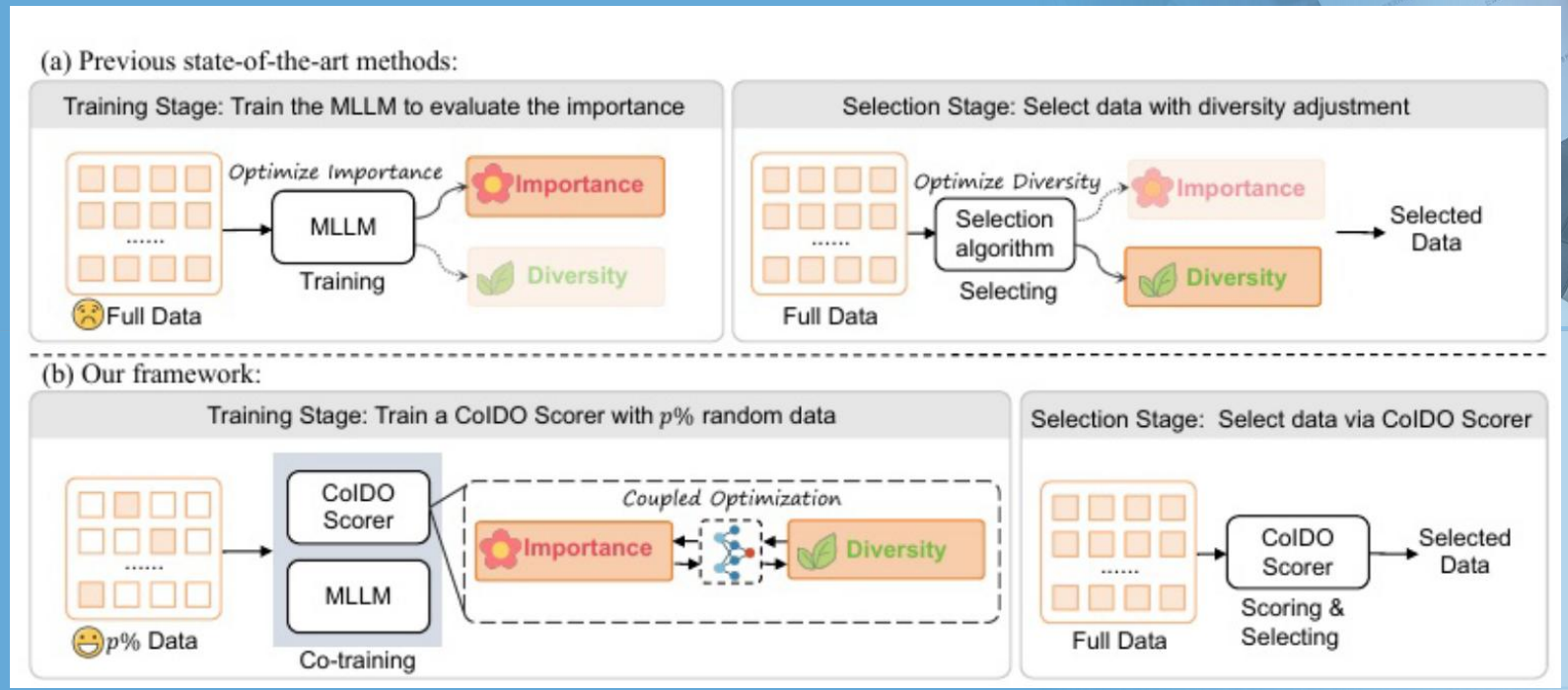
以前最先进的方法(TIVE和ICONS)将重要性和多样性视为解耦且独立的组件,在训练阶段使用整个数据。我们的COIDO在耦合和互反优化中集成了重要性和多样性,通过仅利用整个数据集的一小部分p%(p≪100)进行模型训练,并且在选择阶段没有任何专门的算法,实现了卓越的数据选择。



Background

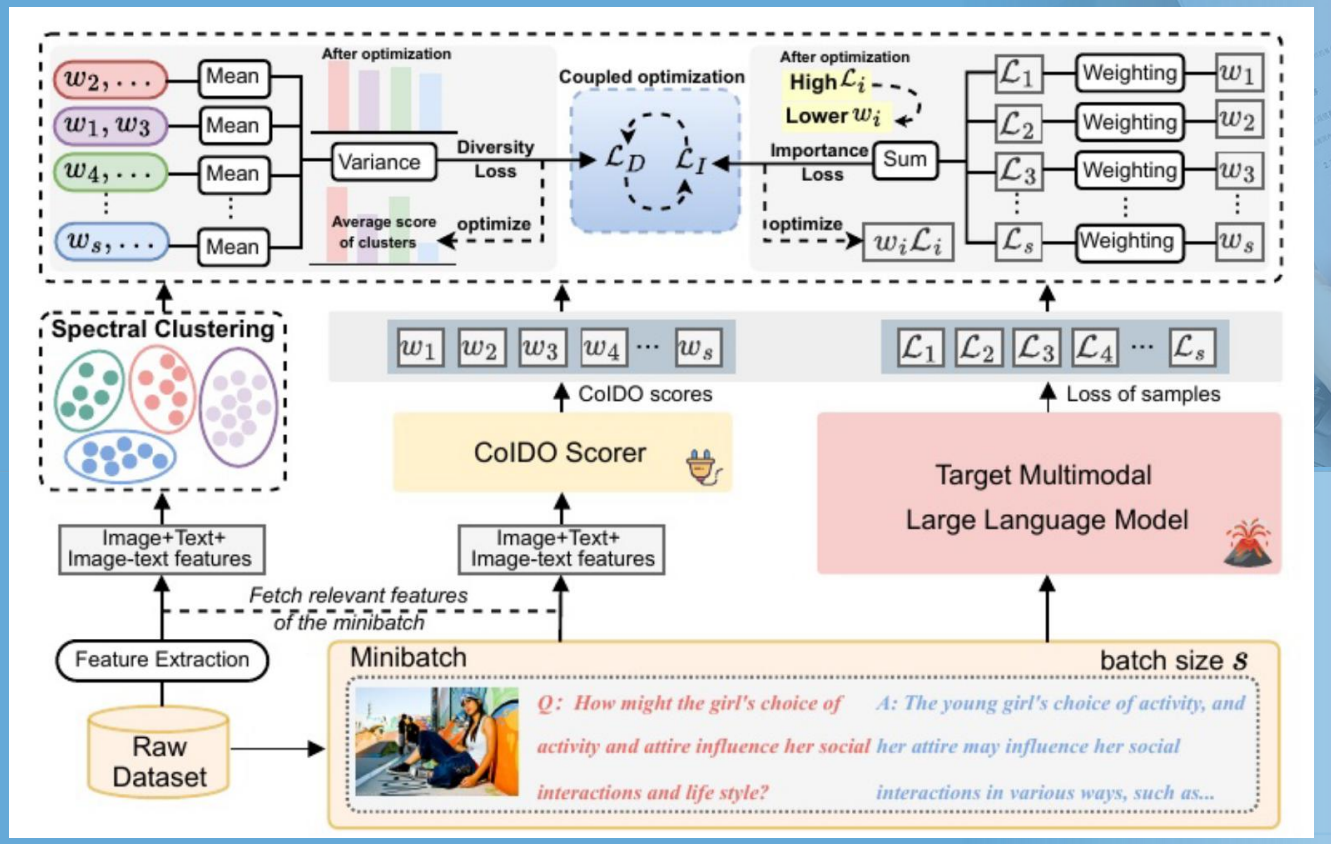
COIDO,一个新的双目标框架,共同优化数据的重要性和多样性,以克服这些挑战。不像现有的方法需要对整个数据集进行昂贵的评估, COIDO使用了一个轻量级的插件评分器。

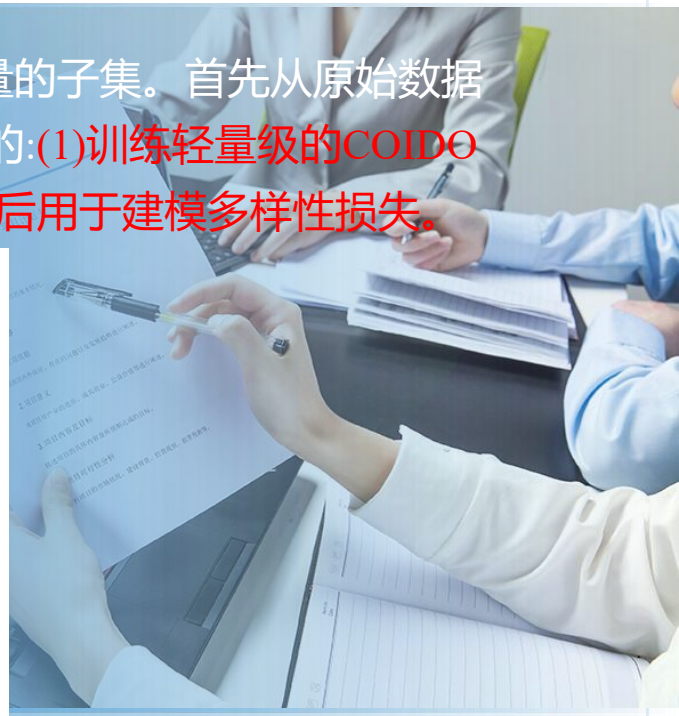
这个计分器只在一个小的随机数据子集上进行训练,以学习候选集的分布,大大降低了计算需求。通过利用基于均方差不确定性的公式,COIDO在训练过程中有效地平衡了重要性和多样性,能够推断所有样本的COIDO分数。这种统一的评分方法允许直接对最有价值的子集进行排名和选择,完全避免了对专门算法的需要。



Method

COIDO框架通过仅利用一小部分p%(例如,20%)的随机样本进行训练, 有效地选择高质量的子集。首先从原始数据集中提取多模态特征和评分指标 (评估文本、图像、图像-文本对齐)。这些特征用于两个目的:(1)训练轻量级的COIDO Scorer来评估一次通过的数据重要性和多样性 (2)聚类以获得每个数据样本的类分配,这将随后用于建模多样性损失。

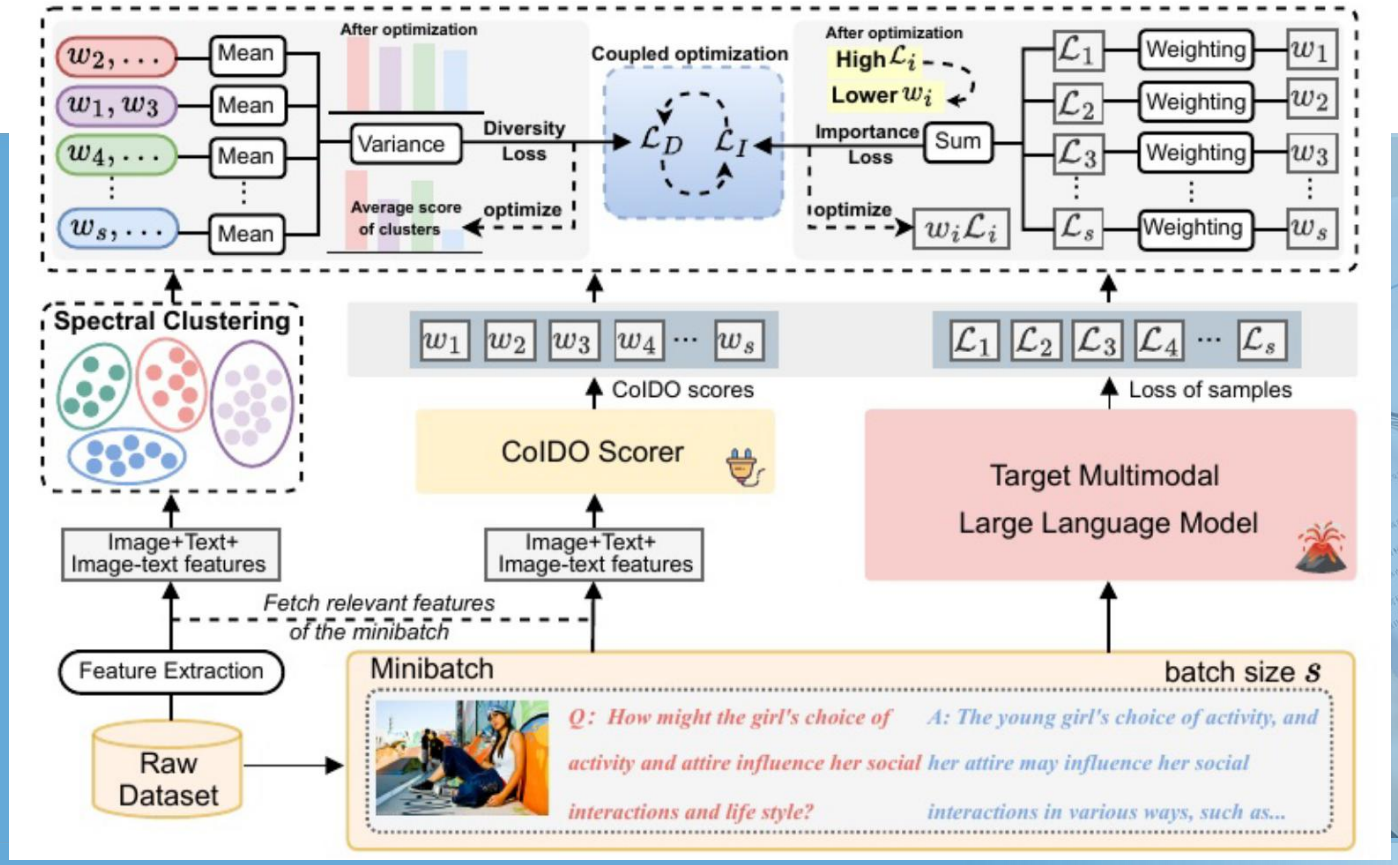




Method

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

重要性与多样性的耦合优化 (Coupled Optimization)



$$\mathcal{L}_D = \mathrm{Var}(\{\bar{w}_1, \bar{w}_2, \ldots, \bar{w}_m\}), \quad \bar{w}_i = \frac{1}{n_i}\sum_{k=1}^{n_i} w_{ik},$$

多样性损失 (L_D)：基于谱聚类（Spectral Clustering）的方差最小化。我们在特征空间将数据聚类，并计算各簇（Cluster）平均得分的方差。通过最小化该方差，迫使模型在挑选高分样本时，不会过度集中于某一类，从而保证了数据的多样性分布。

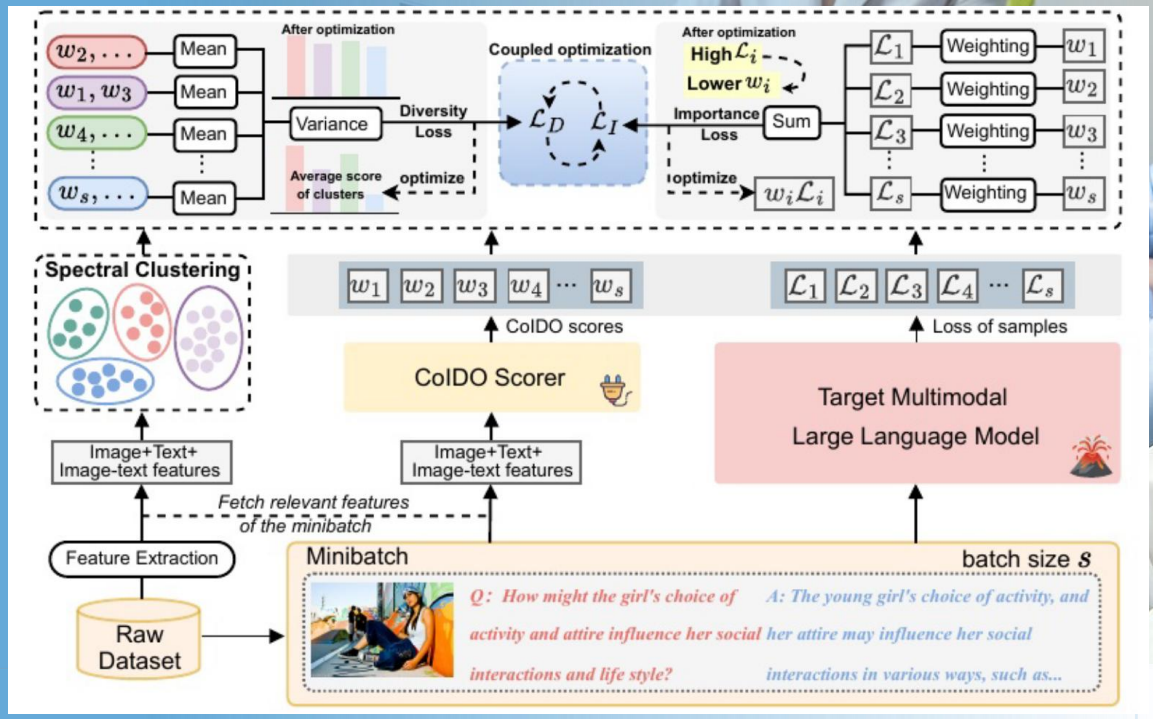
$$\mathcal{L}_I = \sum_{i=1}^{m}\sum_{k=1}^{n_i} w_{ik}\cdot \mathrm{CE}(y_{ik}, \hat{y}_{ik}),$$

重要性损失 (L_I)：基于 Cross-Entropy Loss 的重加权。我们将评分器输出的得分 w 加权作用于 MLLM 的预测 Loss。根据反向传播原理，模型会自动降低高难度（高 Loss）样本的权重以最小化整体 Loss，从而使得评分器隐式地学习到样本的重要性（即：分数越低，样本越重要 / 越难）

Method

基于同方差不确定性的自动加权：为了解决多目标优化中权重超参数难以调节的问题，作者使用任务特定不确定性下的最大似然估计(MLE)框架来表述这个问题。每个损失项被视为概率模型的负对数似然,其中可学习参数捕获优化每个目标时的固有不确定性或噪声。这个概念被称为均方差不确定性,它假设不确定性对于给定目标是恒定的,但在不同目标之间是不同的,这与本文的多模态场景很好地一致。

$$\log p(\mathbf{y}, \mathbf{w} \mid \theta, \sigma_I, \sigma_D) = \sum_{i,k} \log p(y_{ik} \mid x_{ik}, \theta, \sigma_I) + \sum_i \log p(\bar{w}_i \mid \theta, \sigma_D).$$



Method

多样性目标的推导：对于多样性损失 $\mathcal{L}_D$，本文将其建模为一个满足高斯分布的回归问题，假设不同聚类簇的平均权重服从方差为 σ_D^2 的高斯分布。通过最大化该分布的对数似然，导出多样性目标形式。

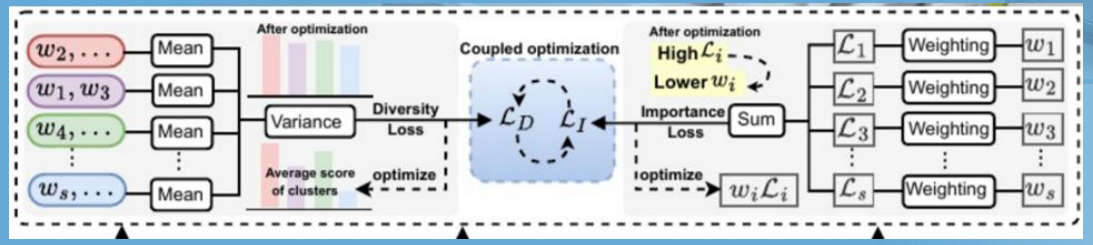
$$\mathcal{L}_D = \mathrm{Var}(\{\bar{w}_1, \bar{w}_2, \ldots, \bar{w}_m\}), \quad \bar{w}_i = \frac{1}{n_i}\sum_{k=1}^{n_i} w_{ik},$$

重要性目标的推导：对于重要性损失 $\mathcal{L}_I$，本文将其构建在一个实例加权的极大似然估计（MLE）框架下，采用加权玻尔兹曼分布（Boltzmann Distribution）来建模样本预测概率。在推导其负对数似然函数时，针对其中的对数配分函数项，本文进行了二阶泰勒展开（Second-order Taylor Expansion）。这一展开过程自然地引入了预测分布的熵（Entropy）H (p)。由于在大模型生成任务中，有效的候选 Token 数量远小于词表大小，根据熵的定义本文能推导出展开式的一阶误差项有一个很小的上界，因此该项可以被忽略。最终，重要性目标被简化为由 σ_I 缩放的交叉熵损失形式。

$$\mathcal{L}_I = \sum_{i=1}^{m}\sum_{k=1}^{n_i} w_{ik} \cdot \mathrm{CE}(y_{ik}, \hat{y}_{ik}),$$

$$\log p(\mathbf{y}, \mathbf{w} \mid \theta, \sigma_I, \sigma_D) = \sum_{i,k} \log p(y_{ik} \mid x_{ik}, \theta, \sigma_I) + \sum_i \log p(\bar{w}_i \mid \theta, \sigma_D).$$

$$p(\bar{w}_i|\theta, \sigma_D) = \mathcal{N}(\bar{w}_i; \mu, \sigma_D^2).$$



$$p(y_{ik} \mid x_{ik}, \theta, \sigma_I, w_{ik}) = \mathrm{Softmax}\left(\frac{w_{ik}}{\sigma_I^2} f_\theta(x_{ik})\right),$$

$$-\sum_i \log p(\bar{w}_i \mid \theta, \sigma_D) = \frac{1}{2\sigma_D^2}\mathcal{L}_D + \log \sigma_D,$$

$$-\sum_{i,k} \log p(y_{ik} \mid x_{ik}, \theta, \sigma_I) = \frac{1}{\sigma_I^2}\mathcal{L}_I + \log \sigma_I.$$

$$\mathcal{L}_{\text{total}} = \frac{1}{\sigma_I^2}\mathcal{L}_I + \frac{1}{2\sigma_D^2}\mathcal{L}_D + \log \sigma_I + \log \sigma_D.$$

目标	概率模型	为什么选择?	数学含义
多样性	高斯分布	多样性是"连续、平滑"的平衡问题；每个簇的平均权重应接近目标值	建模为回归问题，最小化簇间差异
重要性	加权玻尔兹曼分布	重要性是"离散、排序"的选择问题；高分样本应有更高概率被选中	建模为软加权，自动分配样本权重

Experiments

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

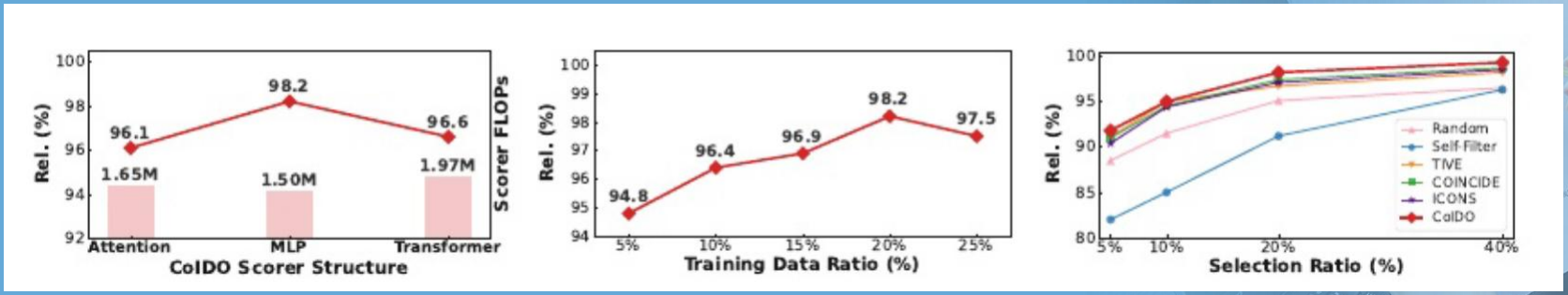
本文在 LLaVA-1.5-7B 模型及 LLaVA-665K 视觉指令调优数据集上进行了广泛验证，并在 10 个主流多模态基准（包括 VQAv2, GQA, MMBench 等）上进行了测试。

Method	VQAv2	GQA	VizWiz	SQA-I	TextVQA	POPE	MME	MMBench en	MMBench cn	LLaVA-Bench	*Rel.* (%)	MLLM Training Data Cost (%)	Total FLOPs
Full Data	79.1	63.0	47.8	68.4	58.2	86.4	1476.9	66.1	58.9	67.9	100	\	10.2E
Model-free Methods													
RANDOM	75.9	59.3	43.6	68.6	55.3	85.9	1461.0	60.3	53.3	64.5	95.1	\	\
CLIP-SCORE [29]	73.4	51.4	43.0	65.0	54.7	85.3	1331.6	55.2	52.0	66.2	91.2	\	\
EL2N [32]	76.2	58.7	43.7	65.5	53.0	84.3	1439.5	53.2	47.4	64.9	92.0	\	\
PERPLEXITY [33]	75.8	57.0	47.8	65.1	52.8	82.6	1341.4	52.0	45.8	68.3	91.6	\	\
SEMDEDUP [30]	74.2	54.5	46.9	65.8	55.5	84.7	1376.9	52.2	48.5	70.0	92.6	\	\
D2-PRUNING [31]	73.0	58.4	41.9	69.3	51.8	85.7	1391.2	65.7	57.6	63.9	94.8	\	\
SELF-SUP [30]	74.9	59.5	46.0	67.8	49.3	83.5	1335.9	61.4	53.8	63.3	93.4	\	\
Model-involved Methods													
SELF-FILTER [20]	73.7	58.3	**53.2**	61.4	52.9	83.8	1306.2	48.8	45.3	64.9	90.9	100	31.2E
TIVE♣♦ [17]	76.0	58.4	44.6	69.8	53.3	85.7	1448.4	**66.9**	**58.7**	63.4	96.7	100+*8*	11.7E
ICONS♣♢ [19]	77.0	**60.4**	45.5	**70.4**	54.5	**86.1**	1447.7	64.6	54.0	66.9	97.1	100+*5*+*2.2*	12.6E
COINCIDE [21]	76.5	59.8	46.8	69.2	**55.6**	**86.1**	**1495.6**	63.1	54.5	67.3	97.4	100	4.9E
CoIDO (Ours)	**77.2**	**60.4**	47.1	69.4	**55.6**	85.4	1450.2	63.8	56.7	**70.1**	**98.2**	**20**	**4.2E**

Loss Function	VQAv2	GQA	Vizwiz	SQA-I	TextVQA	POPE	MME	MMBench(en)	MMBench(cn)	LLAVA-B	*Rel.* (%)
$\mathcal{L}_I$	**77.9**	48.9	44.6	59.7	52.5	**86.2**	1393.5	51.1	44.9	64.9	89.0
$\mathcal{L}_I + \mathcal{L}_D$	74.5	55.8	46.4	67.3	52.6	83.5	1339.7	57.0	50.9	62.3	92.0
$\lambda\mathcal{L}_I + (1-\lambda)\mathcal{L}_D$	76.1	59.4	46.8	68.7	54.4	85.2	**1465.6**	60.5	54.0	64.6	95.9
Ours	77.2	**60.4**	**47.1**	**69.4**	**55.6**	85.4	1450.2	**63.8**	**56.7**	**70.1**	**98.2**

Experiments

本文在 LLaVA-1.5-7B 模型及 LLaVA-665K 视觉指令调优数据集上进行了广泛验证，并在 10 个主流多模态基准（包括 VQAv2, GQA, MMBench 等）上进行了测试。



Model / Setting	VQAv2	GQA	VizWiz	SQA	POPE	TextVQA	MME	MMBench(en)	MMBench(cn)	LLaVA-B	*Rel.* (%)
Full Fine-tune	74.5	47.1	52.8	61.8	46.4	85.7	1480.6	40.2	46.2	38.2	100.0
Random	74.6	44.3	50.0	59.8	40.9	81.3	1407.1	49.2	**48.3**	33.6	97.8
CoIDO	**75.7**	45.1	**53.5**	62.3	**45.3**	82.8	1452.9	**52.0**	46.8	37.6	102.1
CoIDO‡	**75.7**	**46.8**	53.3	**66.2**	42.1	**85.5**	**1486.1**	51.4	47.3	**40.8**	**103.7**

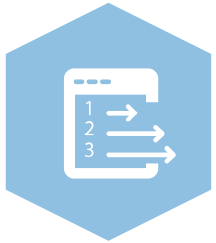
南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

SegEarth-OV: Towards Training-Free Open-Vocabulary Segmentation for Remote Sensing Images

Kaiyu Li[1], Ruixun Liu[2], Xiangyong Cao[2,4†], Xueru Bai[6], Feng Zhou[6], Deyu Meng[3,4,5], Zhi Wang[1]
[1]School of Software Engineering, Xi'an Jiaotong University, Xi'an, 710049, China
[2]School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, 710049, China
[3]School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, 710049, China
[4]Ministry of Education Key Laboratory of Intelligent Networks and Network Security,
Xi'an Jiaotong University, Xi'an, 710049, China [5]Pengcheng Laboratory [6]Xidian University

汇报人：蒋明忠

时间：2025.12



CVPR

Background

遥感图像改变了人类观察和认识地球的方式。它使我们能够监测土地覆盖/利用类型,有效应对自然灾害(如火 灾、地震、洪水),深入了解食物和水资源等。

值得注意的是,遥感数据可以被认为是计算机视觉中一个独特的模态。与自然图像相比,它涉及到更多样化的空间分辨率(从厘米 到公里)、时间维度(从小时到几十年)和物体视角(头顶和定向)。

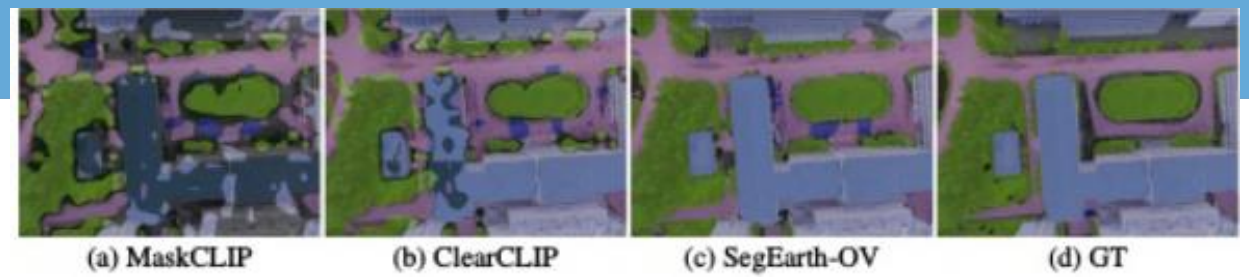
此外,在广阔的地球表面上,“物”(如草地、森林等)比“物” (如建筑物、船舶等)占据的面积要大得多。因此,对于遥感图像而言，语义分割的应用频率要高于实例分割，而对语义级标注的需求加剧了获取大规模标签的难度。目前的遥感语义分割方法大多建立在闭集假设上, 这意味着模型只能识别训练集中存在的预定义类别。然而,在实际的遥感对地观测中,有无数的新类别,手工标注是不切实际的。

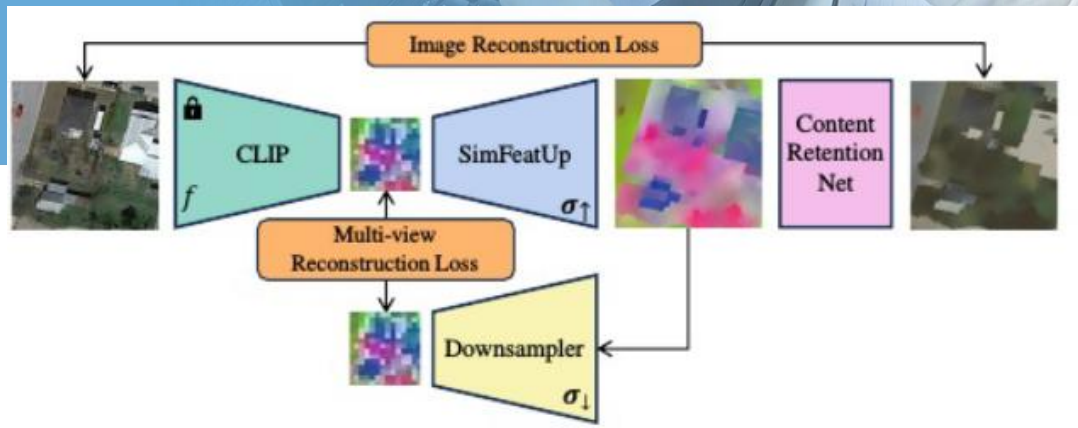
Background

视觉语言模型(VLM)的兴起以其开放词汇语义分割(OVSS)的能力给我们带来了新的启示。然而,通过一些探索性实验，我们发现为自然图像设计的解决方案在遥感图像上是糟糕的。一个值得注意的现象是，预测掩模中存在扭曲的目标形状和不拟合的边界。

根据经验,这些问题在很大程度上可归因于特征的分辨率过低（特别是遥感图像）。在当前基于CLIP的OVSS范例中, 来自CLIP的特征映射被下采样到原始图像的1/16 (vitb /16)。

作者提出了一种简单而通用的上采样器，即simfeature up，以恢复深度特征丢失的空间信息。具体而言，simfeature up只需要从少量未标记的图像中学习，并且可以对任意遥感图像特征进行上采样。

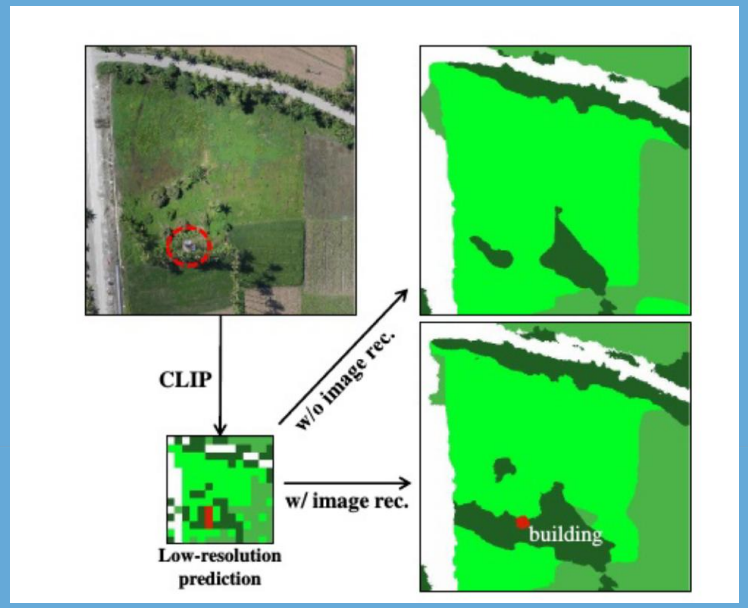




Method

SimFeatUp

feature up为我们提供了一个优秀的通用上采样器训练范例。然而,它缺乏对无训练设置的一些考虑,导致OVSS 任务的次优,特别是在遥感环境中。



$$\mathcal{L}_{rec} = \|\mathcal{O}[1:hw+1] - \sigma_{\downarrow}(\sigma_{\uparrow}(\mathcal{O}[1:hw+1]))\|_2^2.$$

$$\begin{aligned} \boldsymbol{q} = \mathrm{Emb}_q(X), \boldsymbol{k} = \mathrm{Emb}_k(X), \boldsymbol{v} = \mathrm{Emb}_v(X), \\ \boldsymbol{y} = X + \mathrm{SA}(\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}), \\ \boldsymbol{z} = \boldsymbol{y} + \mathrm{FFN}(\mathrm{LN}(\boldsymbol{y})), \end{aligned}$$

$$\overline{\mathcal{O}} = \mathrm{Proj}(\boldsymbol{z}),$$

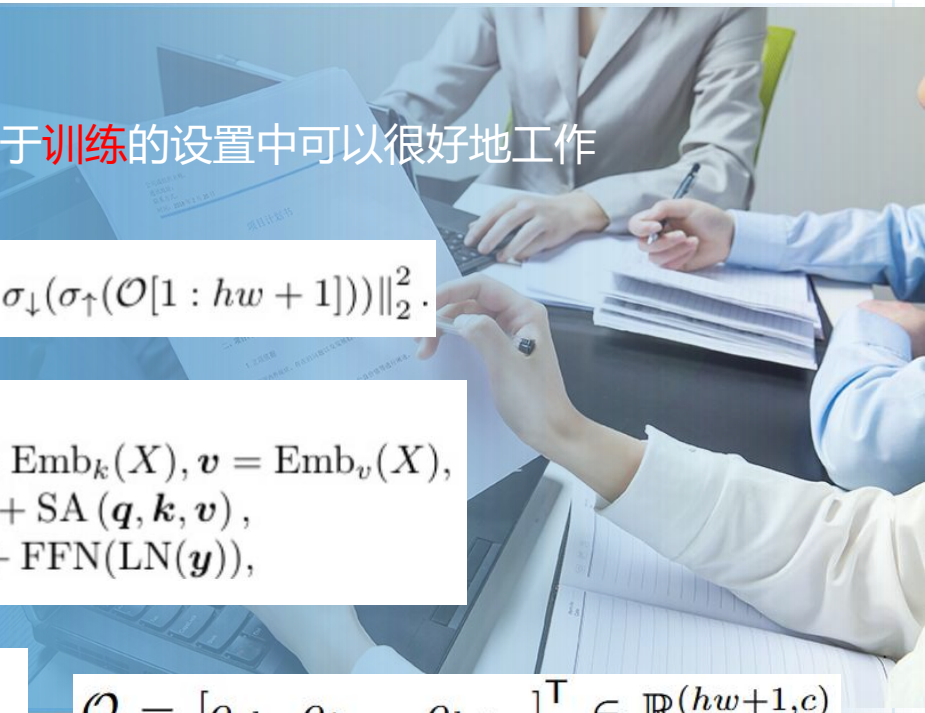
$$\mathcal{O} = [o_{cls}, o_1, \ldots, o_{h\times w}]^{\mathsf{T}} \in \mathbb{R}^{(hw+1,c)}$$

feature up通过对LR特征的相邻元素加权来估计上采样的HR特征元素。 对于权值的生成,JBU考虑两个因素,即制导特征中相邻元素与中心元素之间的相似度和距离,对应于核krange和 kspatail。

Method

SimFeatUp

feature up将CLIP的最终输出,即O [1:hw + 1]作为上采样器的输入。这在基于训练的设置中可以很好地工作



1️⃣ **自注意力的层间演化规律（ViT 已知现象）**

大量 ViT / CLIP 可视化工作表明：

- **浅层**：
  - patch–patch attention 为主
  - 学局部纹理、边缘、低级视觉结构
- **中层**：
  - patch–patch + patch–CLS 混合
- **深层（尤其最后层）**：
  - patch–CLS attention 占主导
  - 表征趋向全局语义一致性

$$\mathcal{L}_{rec} = \|\mathcal{O}[1:hw+1] - \sigma_{\downarrow}(\sigma_{\uparrow}(\mathcal{O}[1:hw+1]))\|_2^2.$$

$$\begin{aligned} \boldsymbol{q} = \mathrm{Emb}_q(X), \boldsymbol{k} = \mathrm{Emb}_k(X), \boldsymbol{v} = \mathrm{Emb}_v(X), \\ \boldsymbol{y} = X + \mathrm{SA}(\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}), \\ \boldsymbol{z} = \boldsymbol{y} + \mathrm{FFN}(\mathrm{LN}(\boldsymbol{y})), \end{aligned}$$

$$\mathcal{O} = \mathrm{Proj}(\boldsymbol{z}),$$

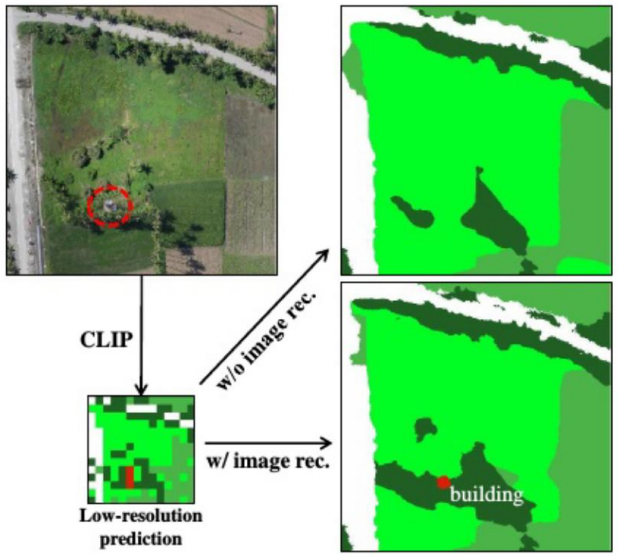
$$\mathcal{O} = [o_{cls}, o_1, \ldots, o_{h\times w}]^{\mathsf{T}} \in \mathbb{R}^{(hw+1,c)}$$

$$\mathcal{O}' = \mathrm{Proj}(X[1:hw+1]).$$

Method

SimFeatUp

为了解决这个问题,作者引入了一个额外的图像重建损失来约束HR特征



$$\mathcal{L}_{rec} = \|\mathcal{O}[1:hw+1] - \sigma_{\downarrow}(\sigma_{\uparrow}(\mathcal{O}[1:hw+1]))\|_2^2.$$

$$\begin{aligned} \boldsymbol{q} = \mathrm{Emb}_q(X), \boldsymbol{k} = \mathrm{Emb}_k(X), \boldsymbol{v} = \mathrm{Emb}_v(X), \\ \boldsymbol{y} = X + \mathrm{SA}(\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}), \\ \boldsymbol{z} = \boldsymbol{y} + \mathrm{FFN}(\mathrm{LN}(\boldsymbol{y})), \end{aligned}$$

$$\mathcal{O} = \mathrm{Proj}(\boldsymbol{z}),$$

$$\mathcal{O} = [o_{cls}, o_1, \ldots, o_{h\times w}]^{\mathsf{T}} \in \mathbb{R}^{(hw+1,c)}$$

$$\mathcal{O}' = \mathrm{Proj}(X[1:hw+1]).$$

$$\mathcal{L}_{img} = \|I - \mathrm{CRN}(\sigma_{\uparrow}(\mathcal{O}[1:hw+1]))\|_2^2,$$

$$\mathcal{L} = \mathcal{L}_{rec} + \gamma\mathcal{L}_{img}.$$

Method

SimFeatUp

遵循Feature up中的上采样运算符,即参数化的JBU。JBU的上采样核 krange和kspatail是从制导特征中一个窗口内的元素计算出 来的

$$k_{spatial}(p,q) = \exp\left(\frac{-\|p-q\|_2^2}{2\tau_{spatial}^2}\right),$$

$$k_{range}(p,q) = \operatorname{softmax}_{(a,b)\in\Omega}\left(\frac{1}{\tau_{range}^2} MLP(G[i,j]) \cdot MLP(G[a,b])\right)$$

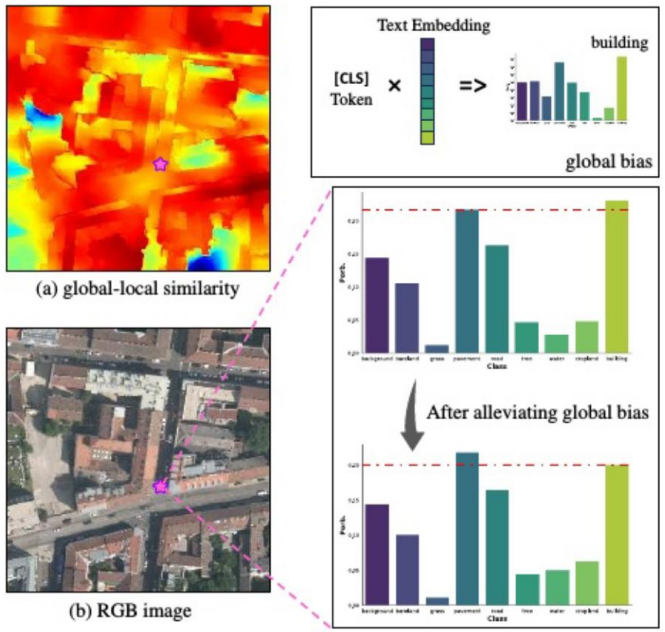
与自然图像不同,目标的大小呈现从米尺度(如树木、花园)到公里尺度(如森林、牧场)的对数尺度。因此,我们设置了更大的上采样核（11×11）。

我们简化了FeatUp中的组件。在 FeatUp中,参数化的JBU模块被堆叠4次进行16倍上采样, 并且每个JBU模块的参数是独立的。虽然我们将HR特征输入到CRN中以确保其内容的完整性,但每个JBU模块的行为是不确定的 。 因 此 , 在 SimFeatUp 中 , 我 们 将 “JBU堆栈”更改为“JBU One”,即只有一个参数化的 JBU用于上采样。

Method

SimFeatUp

CLIP中的每个patch Token都关注于广泛的位置,并且注意图通常具有相似的模式。这表明全局属性CLS Token被附加到CLIP中的patch Token上。这个属性在分类任务中通常不受关注,但它在密集预测中会显著削弱性能.



$$\boldsymbol{q} = \mathrm{Emb}_q(X), \boldsymbol{k} = \mathrm{Emb}_k(X), \boldsymbol{v} = \mathrm{Emb}_v(X),$$
$$\boldsymbol{y} = X + \mathrm{SA}(\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}),$$
$$\boldsymbol{z} = \boldsymbol{y} + \mathrm{FFN}(\mathrm{LN}(\boldsymbol{y})),$$

$$\overline{\mathcal{O}} = \mathrm{Proj}(\boldsymbol{z}),$$

$$\mathcal{O} = [o_{cls}, o_1, \ldots, o_{h\times w}]^{\mathsf{T}} \in \mathbb{R}^{(hw+1,c)}$$

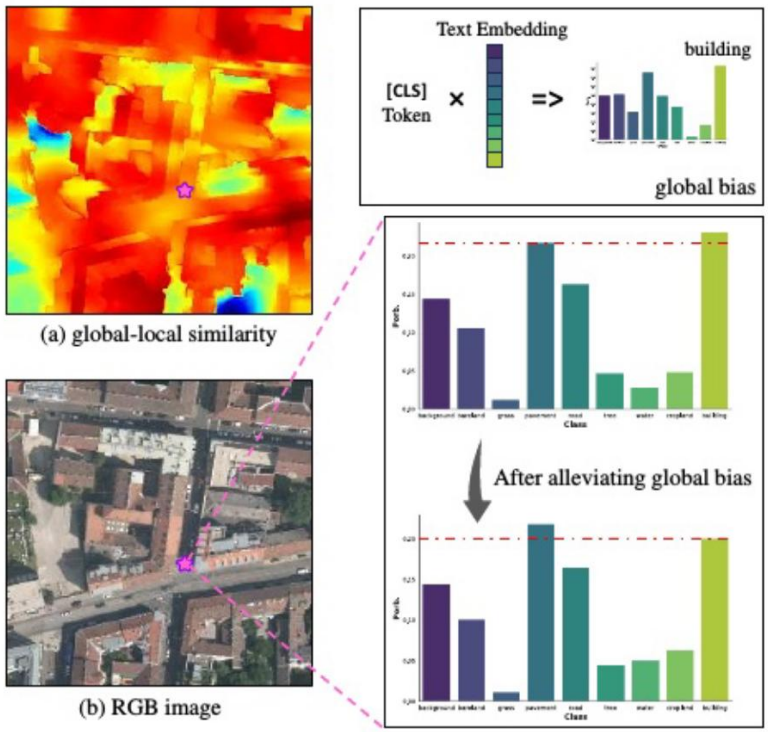
$$\mathcal{O}' = \mathrm{Proj}(X[1:hw+1]).$$

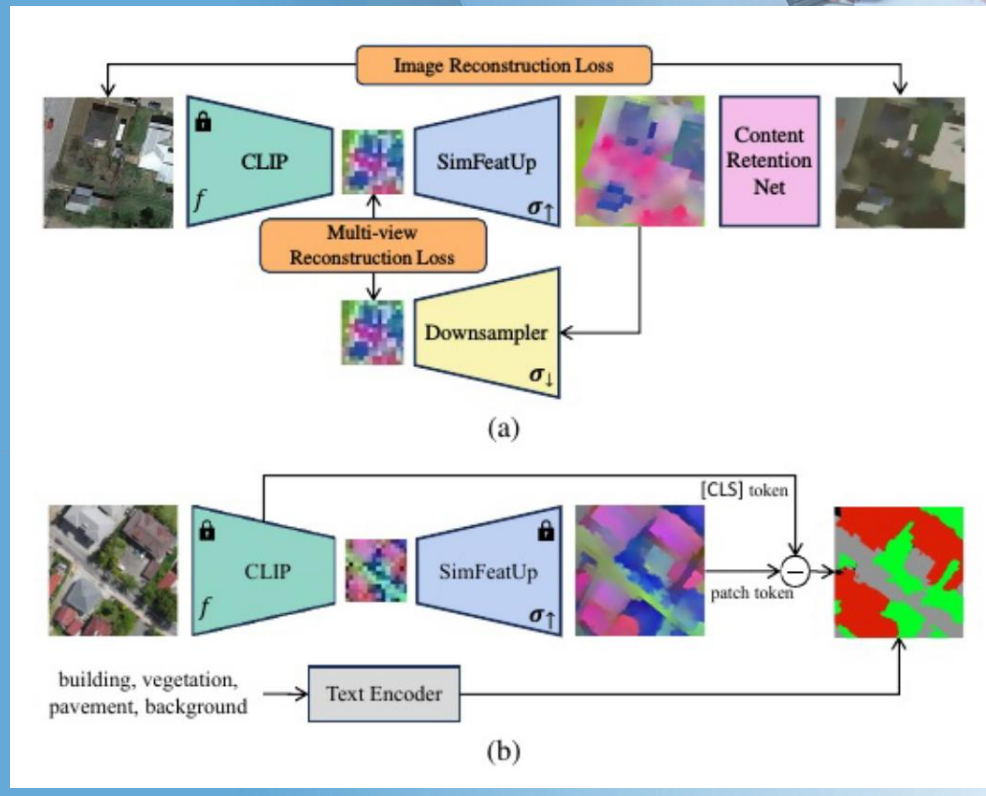
$$\hat{\mathcal{O}} = \mathcal{O}[1:hw+1] - \lambda\mathcal{O}[0],$$

Method

SimFeatUp

CLIP中的每个patch Token都关注于广泛的位置,并且注意图通常具有相似的模式。这表明全局属性CLS Token被附加到CLIP中的patch Token上。这个属性在分类任务中通常不受关注,但它在密集预测中会显著削弱性能.

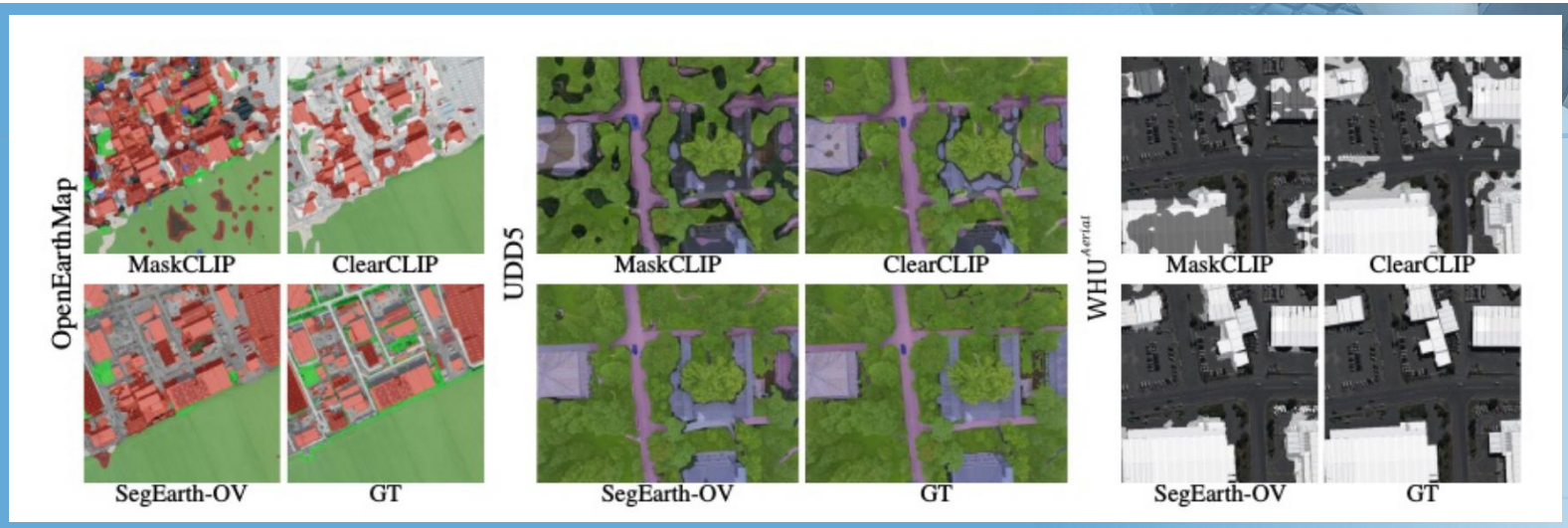




Experiments

在遥感应用环境中,不仅需要多类语义分割（mIoU）,还需要提取某些土地覆盖类型(IoU)(如建筑物、道路、水体),本文选择了17个典型的数据集,涵盖了常见的语义分割、建筑提取、道路提取 和水体分割(洪水检测)任务。

Methods		OpenEarthMap	LoveDA	iSAID	Potsdam	Vaihingen	UAVid^{img}	UDD5	VDD	Average
CLIP [51]	ICML'21	12.0	12.4	7.5	14.5	10.3	10.9	9.5	14.2	11.4
MaskCLIP [77]	ECCV'22	25.1	27.8	14.5	31.7	24.7	28.6	32.4	32.9	27.2
SCLIP [62]	arXiv'23	29.3	30.4	16.1	36.6	**28.4**	31.4	38.7	37.9	31.1
GEM [3]	CVPR'24	33.9	31.6	17.7	36.5	24.7	33.4	41.2	**39.5**	32.3
ClearCLIP [30]	ECCV'24	**31.0**	**32.4**	**18.2**	**40.9**	27.3	**36.2**	**41.8**	39.3	**33.4**
SegEarth-OV	Ours	**40.3**	**36.9**	**21.7**	**47.1**	**29.1**	**42.5**	**50.6**	**45.3**	**39.2**



Experiments

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

在遥感应用环境中,不仅需要多类语义分割（mIoU）,还需要提取某些土地覆盖类型(IoU)(如建筑物、道路、水体),本文选择了17个典型的数据集,涵盖了常见的语义分割、建筑提取、道路提取 和水体分割(洪水检测)任务。

Method	Building Extraction				Road Extraction				Flood Detection
	WHU^{Aerial}	$WHU^{Sat.II}$	Inria	xBD^{pre}	CHN6-CUG	DeepGlobe	Massachusetts	SpaceNet	WBS-SI
448 × 448:									
CLIP [51]	17.7	3.5	19.6	16.0	7.7	3.9	4.9	7.1	18.6
MaskCLIP [77]	29.8	14.0	33.4	29.2	28.1	13.2	10.6	20.8	39.8
SCLIP [62]	33.4	21.0	34.9	25.9	21.1	7.0	7.4	14.9	32.1
GEM [3]	24.4	13.6	28.5	20.8	13.4	4.7	5.1	11.9	39.5
ClearCLIP [30]	36.6	**20.8**	39.0	30.1	25.5	5.7	6.4	16.3	44.9
SegEarth-OV	**49.2**	**28.4**	**44.6**	**37.0**	**35.4**	**17.8**	**11.5**	**23.8**	**60.2**
896 × 896:									
SegEarth-OV	**49.9**	-	**48.9**	**43.1**	32.8	**20.1**	**17.2**	**29.1**	**57.9**

Methods	OpenEarthMap	WHU^{Aerial}	WBS-SI
MaskCLIP	25.1	29.8	39.8
+ ours	28.4↑3.3	35.4↑5.6	48.8↑9.0
SCLIP	29.3	33.4	32.1
+ ours	34.4↑5.1	39.5↑6.1	53.4↑21.3
ClearCLIP	31.0	36.6	44.9
+ ours	39.1↑8.1	51.1↑14.5	60.4↑15.5

Methods	Context59 [47]	Stuff [4]	Cityscapes [13]	Average
TCL [7]	30.3	19.6	23.1	24.3
Reco [57]	22.3	14.8	21.1	19.4
MaskCLIP	26.4	16.4	12.6	18.5
+ SimFeatUp	28.7	18.0	25.8	24.2 ↑5.7
SCLIP	33.0	21.1	29.1	27.7
+ SimFeatUp	34.1	22.0	30.5	28.9 ↑1.2
ClearCLIP	35.9	23.9	30.0	29.9
+ SimFeatUp	37.5	25.1	30.7	31.1 ↑1.2

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

SegEarth-OV3: Exploring SAM 3 for Open-Vocabulary Semantic Segmentation in Remote Sensing Images

Kaiyu Li[1*], Shengqi Zhang[1*], Yupeng Deng[2], Zhi Wang[1], Deyu Meng[1], Xiangyong Cao[1†]
[1]Xi'an Jiaotong University [2] Chinese Academy of Sciences

汇报人：蒋明忠

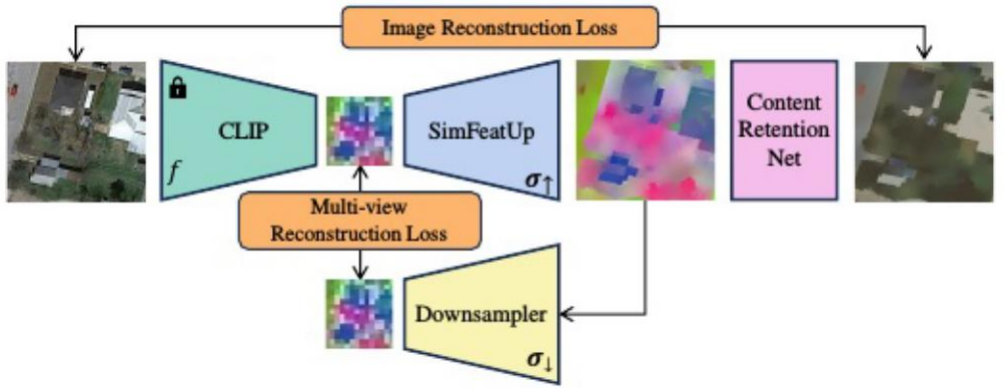
时间：2025.12

Background

现有的无训练开放词汇语义分割(OVSS)方法大多基于 CLIP。虽然这些方法取得了进展,但它们往往面临精确定位的挑战,或者需要复杂的pipelines来组合独立的模块, 特别是在存在大量密集和小目标的遥感场景中更加糟糕。

最近，Segment Anything Model 3 (SAM 3)面世。与基于clip的范例不同，SAM 3是一个支持提示概念分割的统一模型。它建立在DETR和MaskFormer架构之上,采用基于查询的Transformer设计。至关重要的是,SAM 3利用了一种解耦架构,其中Presence Head被专门设计用于预测提示概念在图像中存在的概率。同时,Transformer Decoder (Instance Head)和Semantic Segmentation Head分别为离散实例和连续语义区域生成精确的掩码。

但其对于遥感图像面临着明显的挑战，密集的小物体和巨大的无定形背景的复杂共存。因此,为使SAM 3适应地理空间情景而进行的量身定制探索仍然是有价值的。



Method

本文对SAM 3在遥感OVSS任务中的应用进行了初步探索。在本文中,我们提出了一个初步的探索适应SAM 3 遥感OVSS任务,而不需要额外的训练。我们研究了 SAM 3 的统一架构是否能够提供比复杂的 CLIP-ensemble方法更强大、更简单的baseline。我们提出的方法, 即SegEarth-OV3,由两个适合SAM 3设计的简单策略组成:

Dual-Head Mask Fusion: 首先,我们实现了一个掩码融合策略,该策略结合了SAM 3的语义分割头和Transformer解码器(实例头)的输出。这使我们能够利用两个头的优势来获得更好的鲁棒分割。

Presence-Guided Filtering:其次,我们利用存在头的存在得分来过滤掉场景中不存在的类别,减少遥感场景中庞大的词汇量和Patch级处理造成的误报。

3.1 预备知识：SAM 3 的输出

给定图像 I 和文本提示 t（如 “building”），SAM 3 输出：

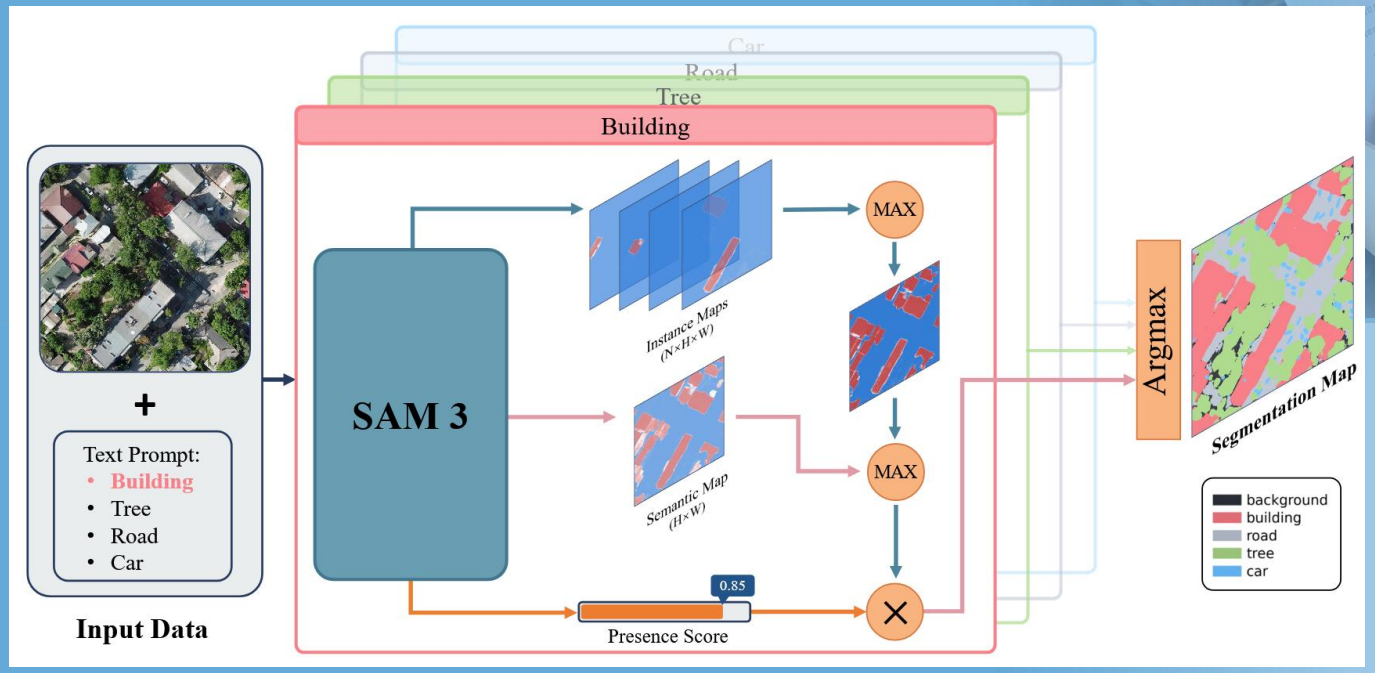
- **存在性分数** $S_{\text{pres}} \in [0,1]$：概念 t 存在于图中的全局概率。
- **语义概率图** $P_{\text{sem}} \in [0,1]^{H \times W}$：来自 FCN 式语义头，保证**区域完整性**。
- **实例预测集** $\{(P_{\text{inst}}^{(k)}, s_{\text{conf}}^{(k)})\}_{k=1}^{N}$：来自 Transformer 解码器，保证**实例边界精度**。

Method

SegEarth-OV3的整体推理流程。给定一个输入图像和一个文本提示列表,我们利用SAM 3的解耦输出。该框架涉及:(1)实例聚合以整合稀疏对象查询;（用实例头）

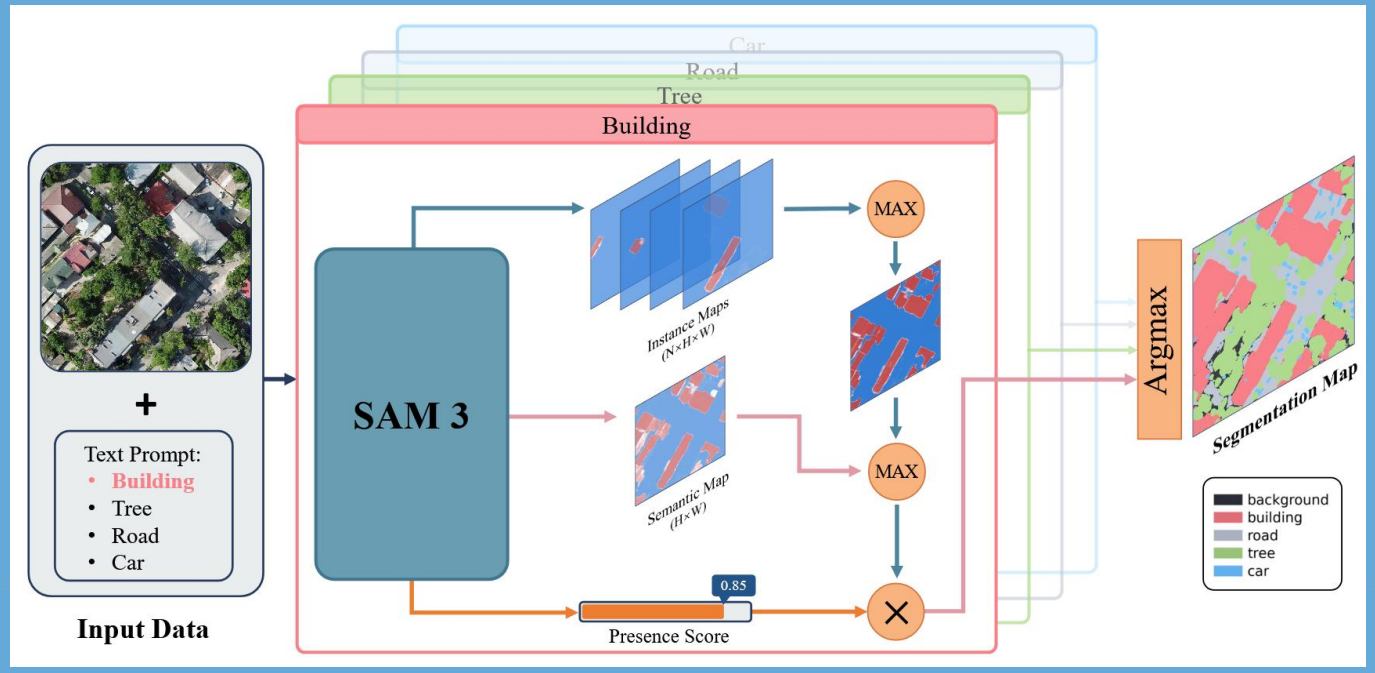
(2)双头掩码融合,将细粒度的实例细节与语义头的全局覆盖相结合;（实例头聚合成特征图与语义头分割的特征图取MAX)

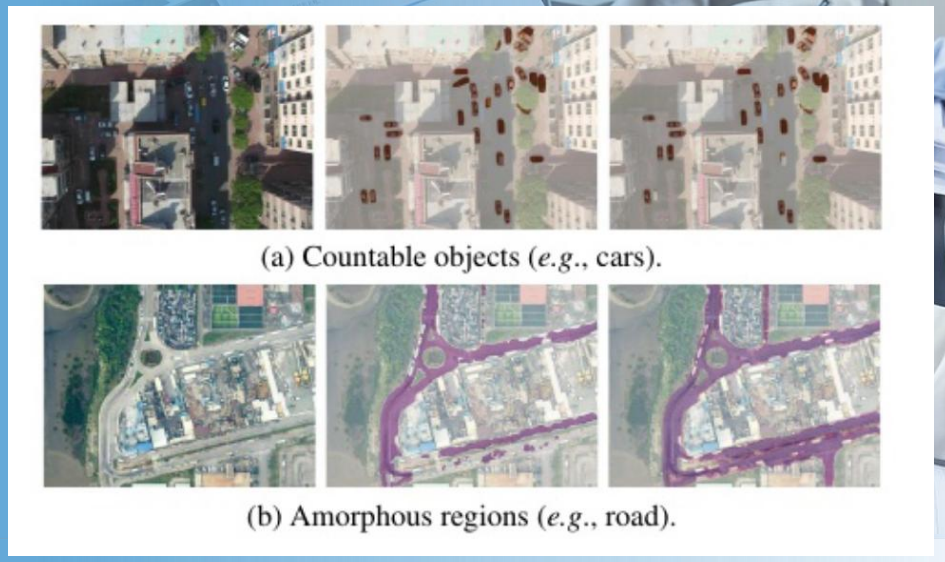
(3)存在引导过滤(使用存在得分),以抑制缺失类别的误报。()



Method

SegEarth-OV3的整体推理流程。给定一个输入图像和一个文本提示列表,我们利用SAM 3的解耦输出。





$$P_{inst\_agg}(h,w) = \max_{k=1}^{N} \left(P_{inst}^{(k)}(h,w) \cdot s_{conf}^{(k)} \right).$$

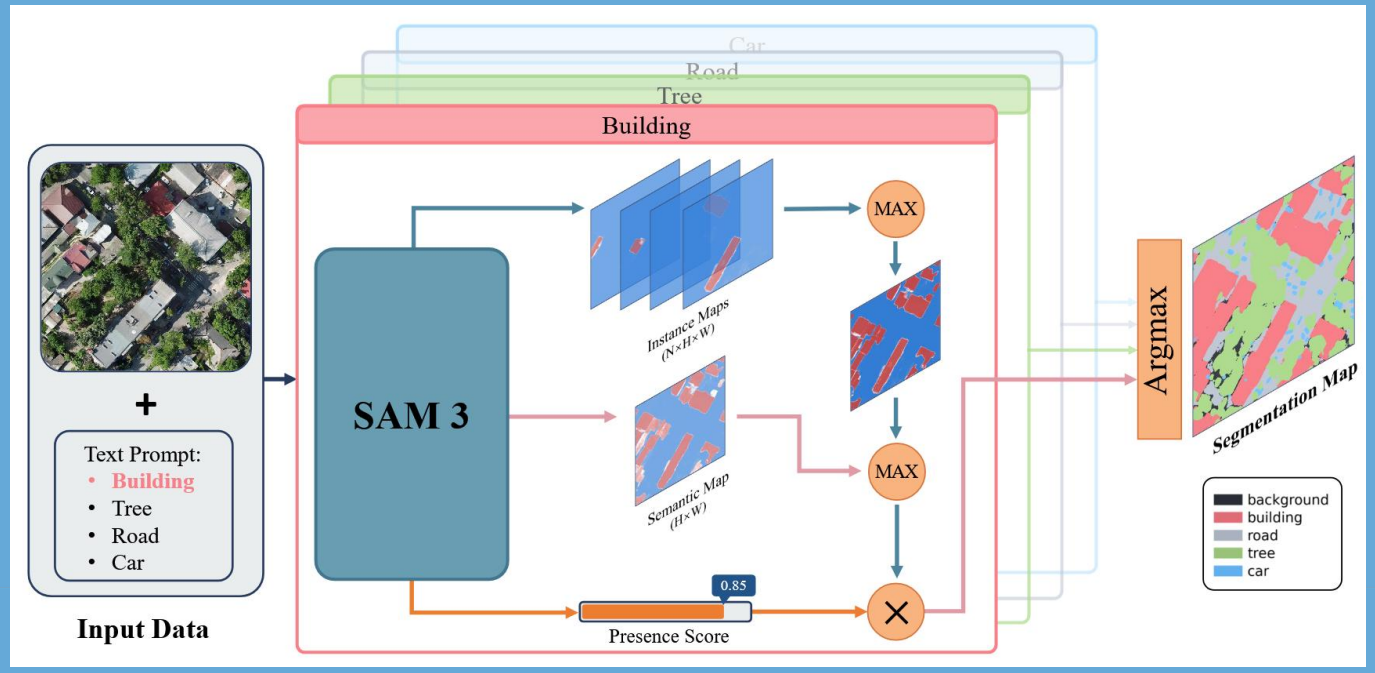
$$P_{fused}(h,w) = \max\left(P_{sem}(h,w), P_{inst\_agg}(h,w)\right).$$

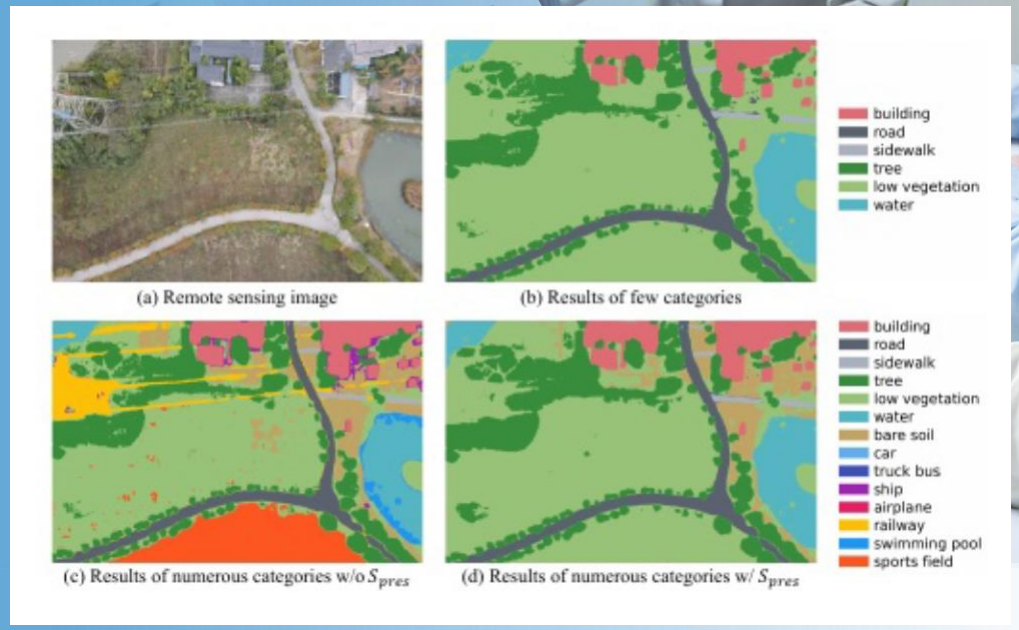
$$P_{final}^{(c)} = P_{fused}^{(c)} \cdot S_{pres}^{(c)}$$

$$M(h,w) = \arg\max_{c \in \mathcal{V}} P_{final}^{(c)}(h,w).$$

Method

SegEarth-OV3的整体推理流程。给定一个输入图像和一个文本提示列表,我们利用SAM 3的解耦输出。





$$P_{inst\_agg}(h,w) = \max_{k=1}^{N} \left(P_{inst}^{(k)}(h,w) \cdot s_{conf}^{(k)} \right).$$

$$P_{fused}(h,w) = \max\left(P_{sem}(h,w), P_{inst\_agg}(h,w)\right).$$

$$P_{final}^{(c)} = P_{fused}^{(c)} \cdot S_{pres}^{(c)}$$

$$M(h,w) = \arg\max_{c \in \mathcal{V}} P_{final}^{(c)}(h,w).$$

Experiments

在SegEarth-OV 之后,作者在17个遥感数据集上评估了我们的方法,涵盖了不同的场景、分辨率和任务。
评估指标为多类分割的mIoU和二元提取任务的前景类的IoU

Methods	OpenEarthMap	LoveDA	iSAID	Potsdam	Vaihingen	UAVid^{img}	UDD5	VDD	Avg
Training on remote sensing segmentation data									
SCAN$_{CVPR2024}$ [42]	-	23.2	44.3	27.5	15.2	20.3	34.1	29.2	-
SAN$_{CVPR2023}$ [65]	-	25.3	49.6	37.3	39.2	23.5	37.2	35.8	-
SED$_{CVPR2024}$ [62]	-	24.6	51.2	29.4	39.0	21.3	35.7	32.5	-
Cat-Seg$_{CVPR2024}$ [18]	-	28.6	53.3	35.8	42.3	25.7	40.2	39.1	-
OVRS$_{TGRS2025}$ [8]	-	31.5	52.7	36.4	43.5	24.1	40.8	37.2	-
GSNet$_{AAAI2025}$ [69]	-	32.5	**53.7**	37.9	44.1	24.2	40.9	37.3	-
RSKT-Seg$_{AAAI2026}$ [36]	-	33.2	**54.3**	38.4	42.7	25.7	42.1	39.7	-
SkySense-O$_{CVPR2025}$ [74]	**40.8**	**38.3**	43.9	**54.1**	**51.6**	-	-	-	-
Training-free									
CLIP$_{ICML2021}$ [48]	12.0	12.4	7.5	15.6	10.8	10.9	9.5	14.2	11.4
MaskCLIP$_{ECCV2022}$ [72]	25.1	27.8	14.5	33.9	29.9	28.6	32.4	32.9	27.2
SCLIP$_{ECCV2024}$ [55]	29.3	30.4	16.1	39.6	35.9	31.4	38.7	37.9	31.1
GEM$_{CVPR2024}$ [5]	33.9	31.6	17.7	39.1	36.4	33.4	41.2	39.5	32.3
ClearCLIP$_{ECCV2024}$ [33]	31.0	32.4	18.2	42.0	36.2	36.2	41.8	39.3	33.4
SegEarth-OV$_{CVPR2025}$ [40]	40.3	36.9	21.7	48.5	40.0	**42.5**	**50.6**	45.3	39.2
ProxyCLIP$_{ECCV2024}$ [34]	38.9	34.3	21.8	49.0	47.5	35.8	40.8	**47.8**	39.5
CorrCLIP$_{ICCV2025}$ [70]	32.9	36.9	25.5	51.9	47.0	38.3	46.1	47.3	**40.7**
SegEarth-OV3	**42.9**	**47.4**	27.6	**57.8**	**60.8**	**54.7**	**71.7**	**64.5**	**53.4**
Oracle	64.4	50.0	36.2	74.3	61.2	59.7	56.5	62.9	58.2

Experiments

在SegEarth-OV 之后,作者在17个遥感数据集上评估了我们的方法,涵盖了不同的场景、分辨率和任务。
评估指标为多类分割的mIoU和二元提取任务的前景类的IoU

Method	Building Extraction				Road Extraction				Flood Detection
	WHU^{Aerial}	$WHU^{Sat.II}$	Inria	xBD^{pre}	CHN6-CUG	DeepGlobe	Massachusetts	SpaceNet	WBS-SI
CLIP [48]	17.7	3.5	19.6	16.0	7.7	3.9	4.9	7.1	18.6
MaskCLIP [72]	29.8	14.0	33.4	29.2	28.1	13.2	10.6	20.8	39.8
SCLIP [55]	33.4	21.0	34.9	25.9	21.1	7.0	7.4	14.9	32.1
GEM [5]	24.4	13.6	28.5	20.8	13.4	4.7	5.1	11.9	39.5
ClearCLIP [33]	36.6	20.8	39.0	30.1	25.5	5.7	6.4	16.3	44.9
SegEarth-OV [40]	**49.2**	**28.4**	**44.6**	**37.0**	**35.4**	**17.8**	**11.5**	**23.8**	**60.2**
SegEarth-OV3	**86.9**	**44.2**	**72.4**	**64.3**	**49.6**	**39.3**	**27.7**	**35.6**	**75.6**

Method	LoveDA	Uavid	xBD^{pre}	CHN6-CUG
Instance Only	32.2	50.4	61.4	38.4
Semantic Only	35.4	47.1	44.9	39.5
SegEarth-OV3	**47.4**	**54.7**	**64.3**	**49.6**

Experiments

作者在三个标准的通用场景数据集上对SegEarth-OV3进 行基准测试

Method	Size	VOC20	Stuff	City
Training-based				
TCL [12]	ViT-B/16	83.2	22.4	24.0
CLIP-DINOiser [60]		80.9	24.6	31.7
CoDe [59]		-	23.9	28.9
CAT-Seg [18]		94.6	-	-
Training-free				
CLIP =[48]	ViT-B/16	41.9	4.4	5.0
MaskCLIP [72]		74.9	16.4	12.6
ClearCLIP [33]		80.9	23.9	30.0
SCLIP [55]		80.4	22.4	32.2
ProxyCLIP [34]		80.3	26.5	38.1
LaVG [30]		82.5	23.2	26.2
CLIPtrase [50]		81.2	24.1	-
NACLIP [26]		83.0	25.7	38.3
Trident [51]		84.5	28.3	42.9
ResCLIP [67]		86.0	24.7	35.9
SC-CLIP [2]		84.3	26.6	41.0
CLIPer [52]		85.2	27.5	-
CASS [31]		87.8	26.7	39.4
CorrCLIP [70]		88.8	31.6	49.4
FreeDA [3]	ViT-L/14	87.9	28.8	36.7
CaR [53]		91.4	-	-
ProxyCLIP [34]		83.2	25.6	40.1
ResCLIP [67]		85.5	23.4	33.7
SC-CLIP [2]		88.3	26.9	41.3
CLIPer [52]		90.0	28.7	-
CorrCLIP [70]		91.5	34.0	51.1
ProxyCLIP [34]	ViT-H/14	83.3	26.8	42.0
Trident [51]		88.7	28.6	47.6
CorrCLIP [70]		91.8	32.7	49.9
SegEarth-OV3	PE-L+/14	**96.8**	**42.8**	**69.7**

Thanks