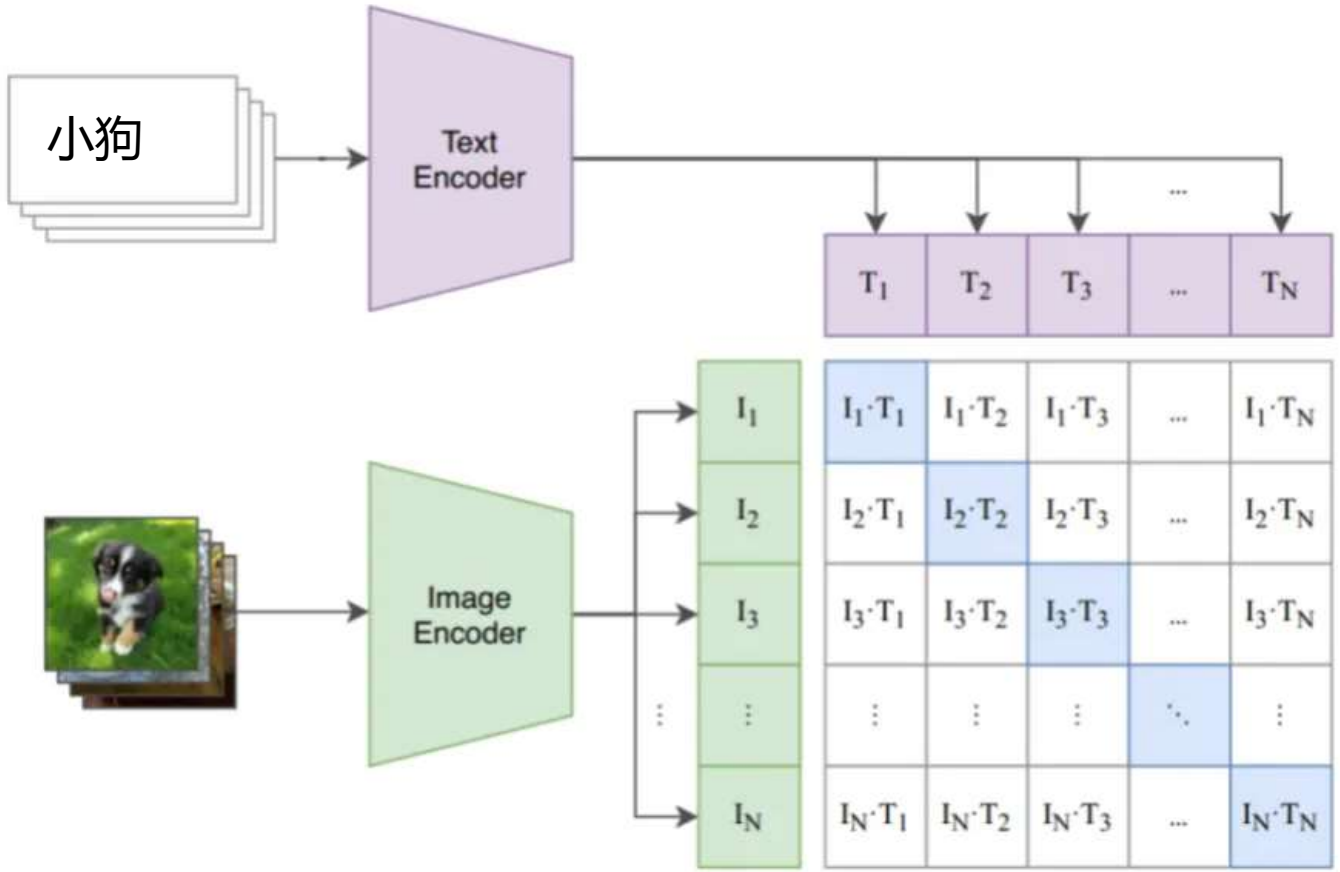
CLIP—Based Change Detection

目录

CLIP(Contrastive Language-Image Pre-Training)

CLIP：利用文本的监督信号训练一个迁移能力强的视觉模型

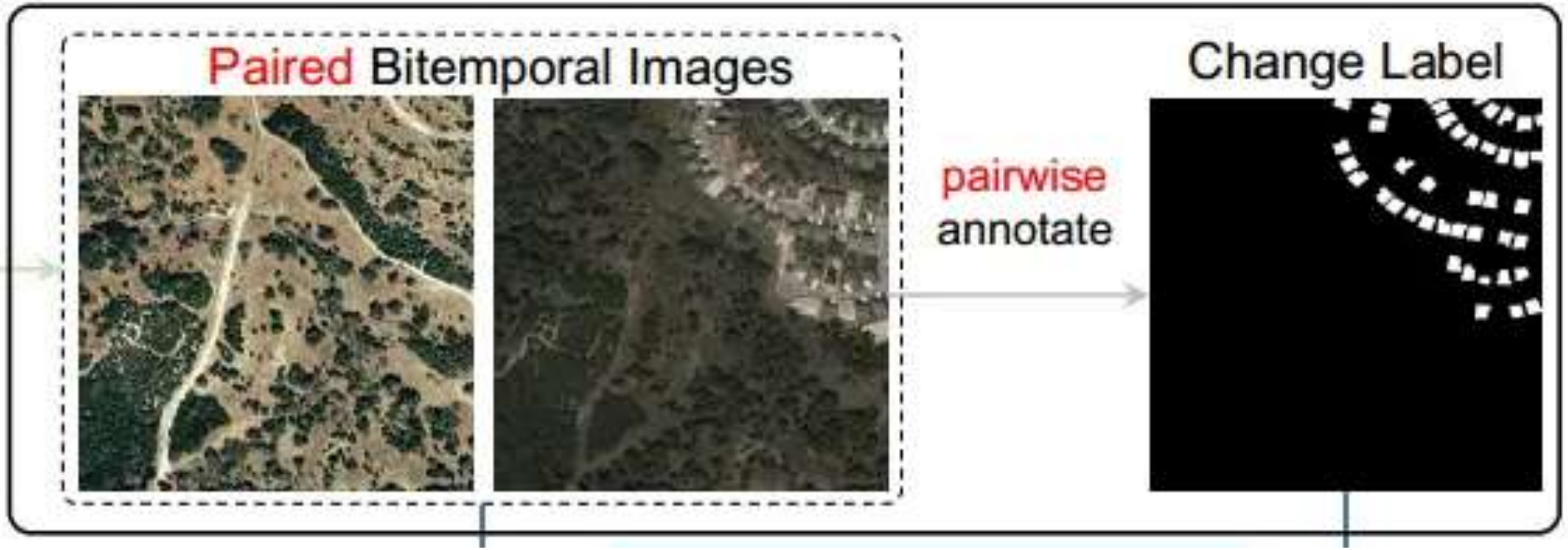


图像-文本匹配

余弦相似性

CLIP原理图

变化检测的目标



$X_{t1,pos}$

$\xrightarrow{F_{IsChanged?}(X_{t1,pos}, X_{t2,pos}) = Y, Y \in \{0,1\}}$ $X_{t2,pos}$

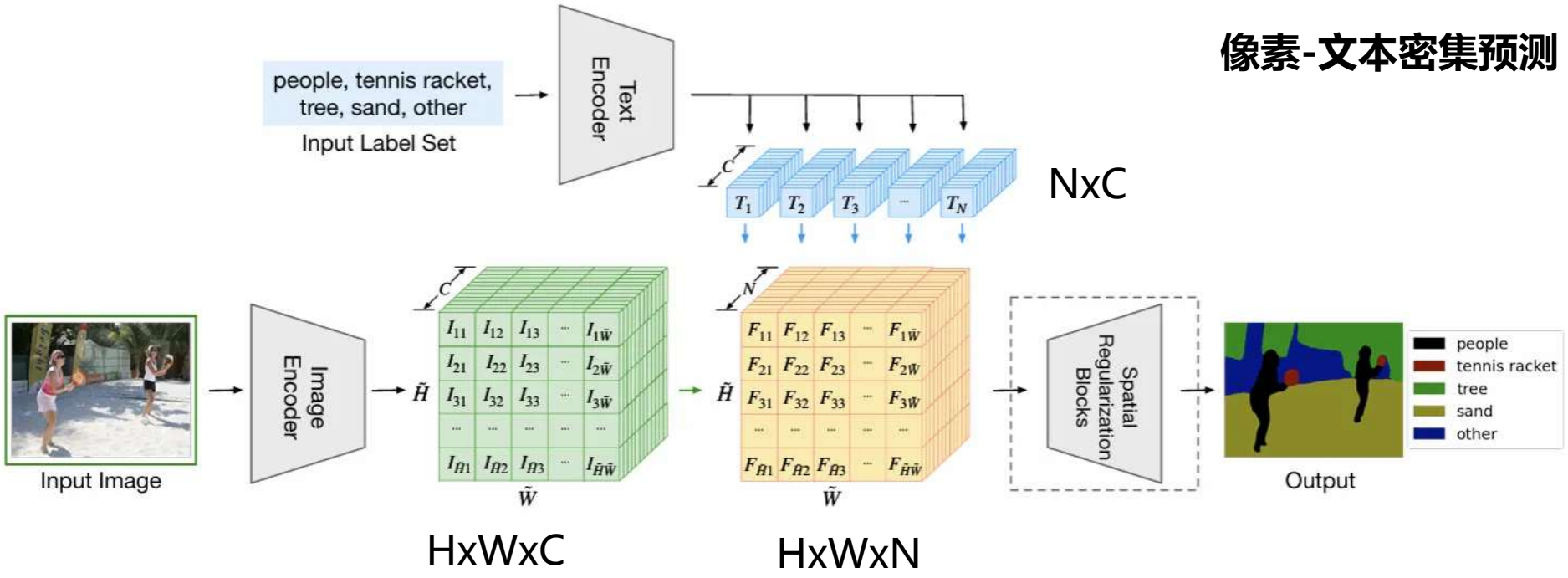
$F_{ChangeToWhat?}(X_{t1,pos}, X_{t2,pos}) = Y, Y \in \{label_1, label_2, \ldots, label_n\}$

Language-Driven Semantic Segmentation

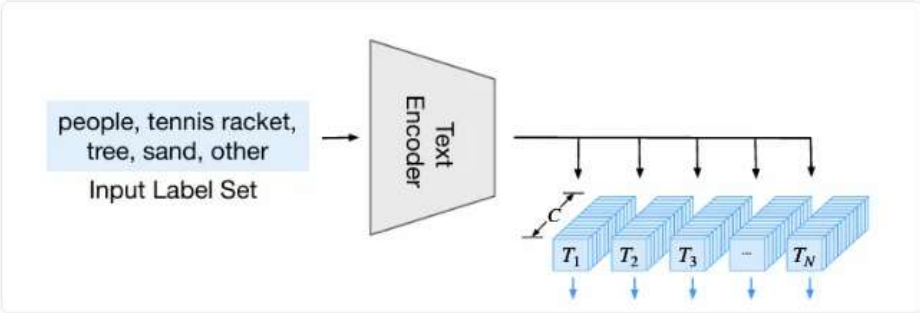
语言驱动的语义分割

研究目标：利用现代语言模型来增加语义分割模型的灵活性和通用性。

语言驱动的语义分割(LSeg)，将文本标签和图像像素嵌入(embedding)到一个公共空间(Tensor, Vector)中，并根据内积为每个像素分配最接近的标签。



文本编码器(Text Encoder)与图像编码器(Image Encoder)

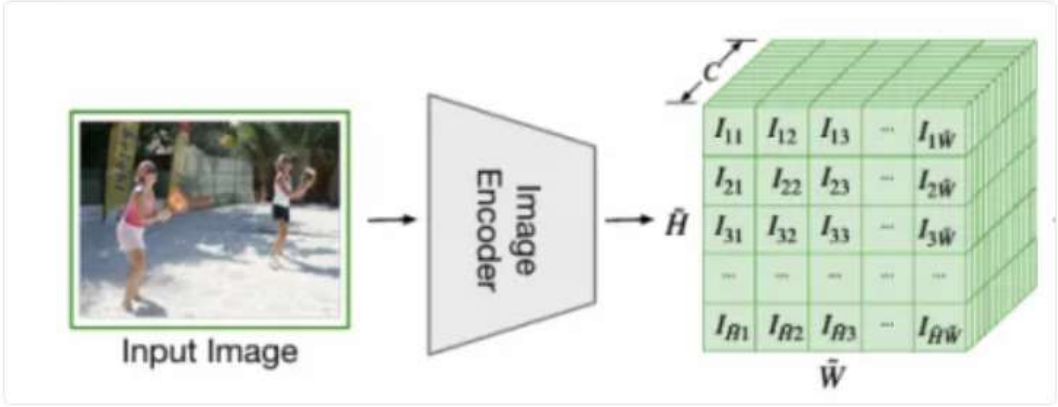


文本编码器(Text Encoder)

标签嵌入(embedding)到嵌入空间

$$T_i = TextEncoder(Text_{input})$$

$$T_{people} = TextEncoder(people) = (T_{11}, T_{12}, \ldots, T_{1c})$$



图像编码器(Image Encoder)

标准化Normalize

先用下采样缩放图像

$$\tilde{H} = \frac{H}{s}, \tilde{W} = \frac{W}{s}$$

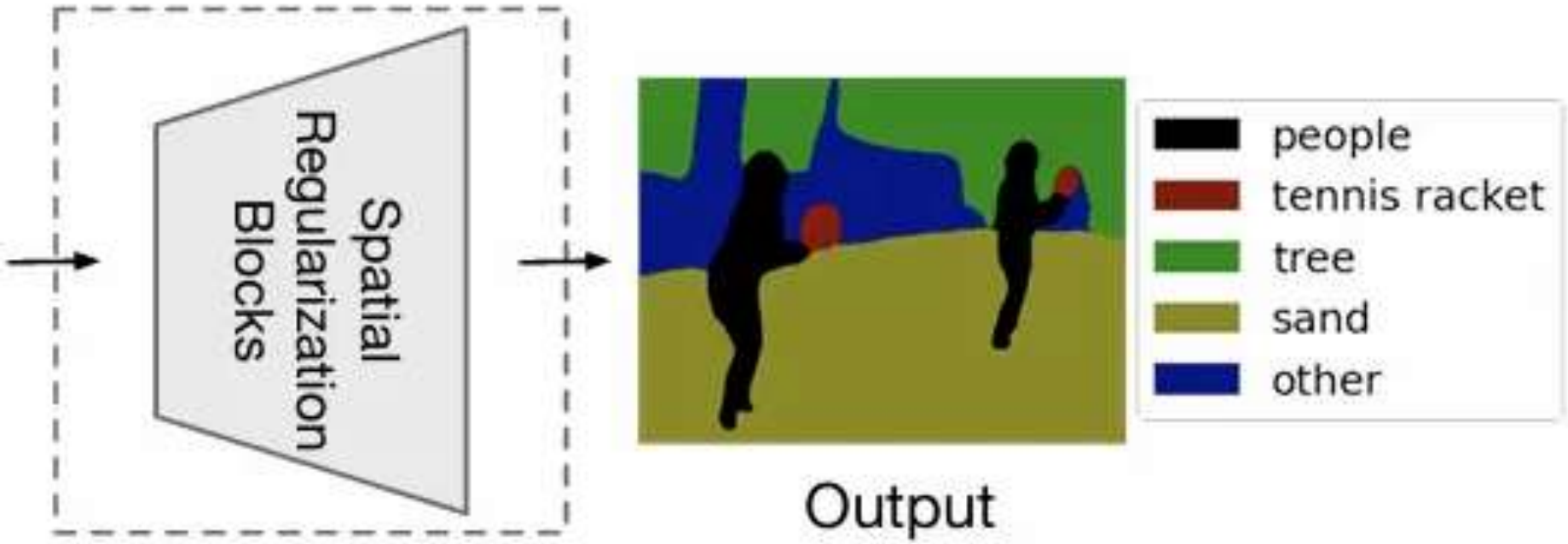
$$Image_{output} = DownSample_{(s)}(Image_{input})$$

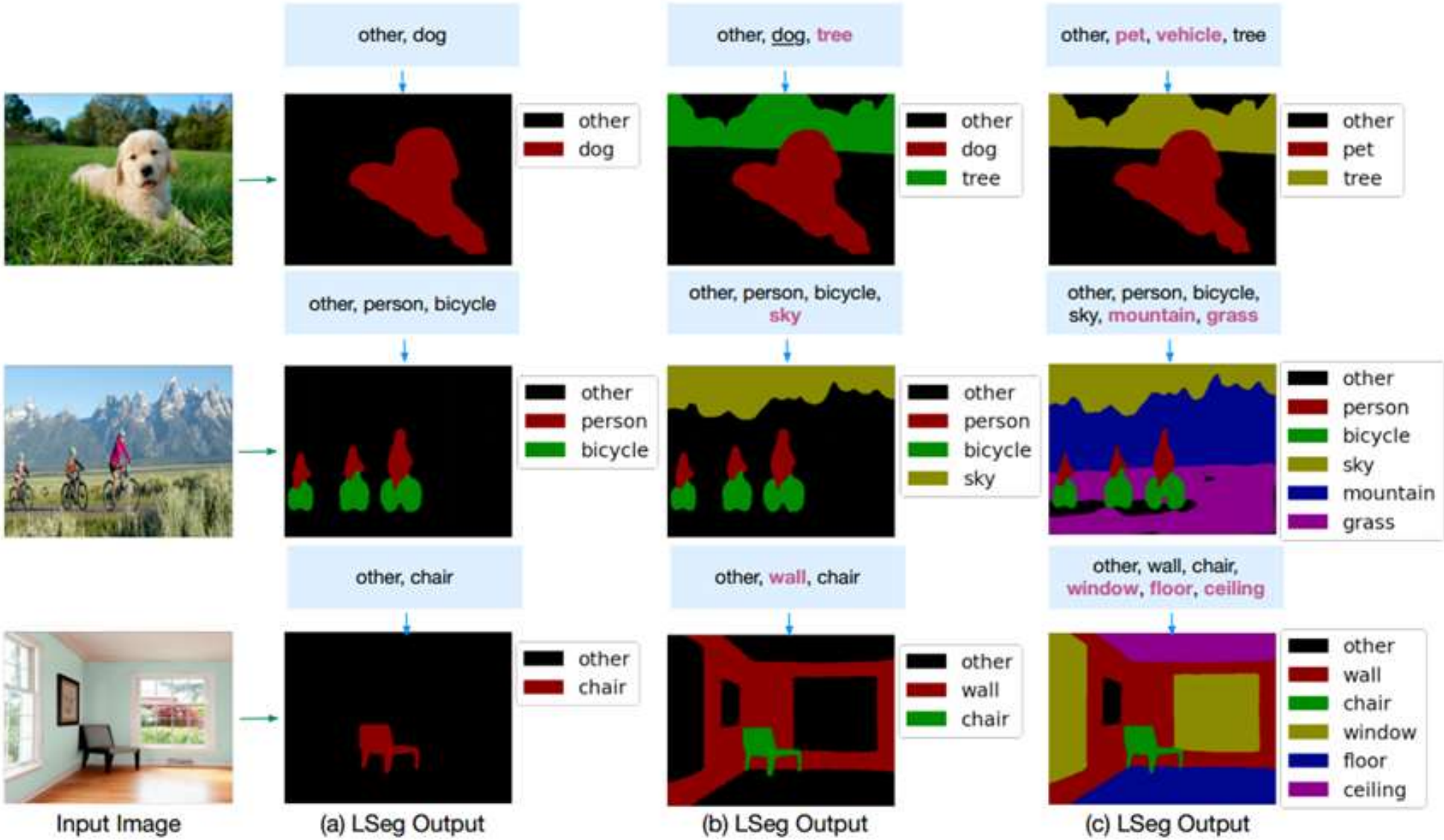
$$Image_{Encoded} = ImageEncoder(Image_{output})$$

空间正则化

文中评估了两种满足此属性的函数：

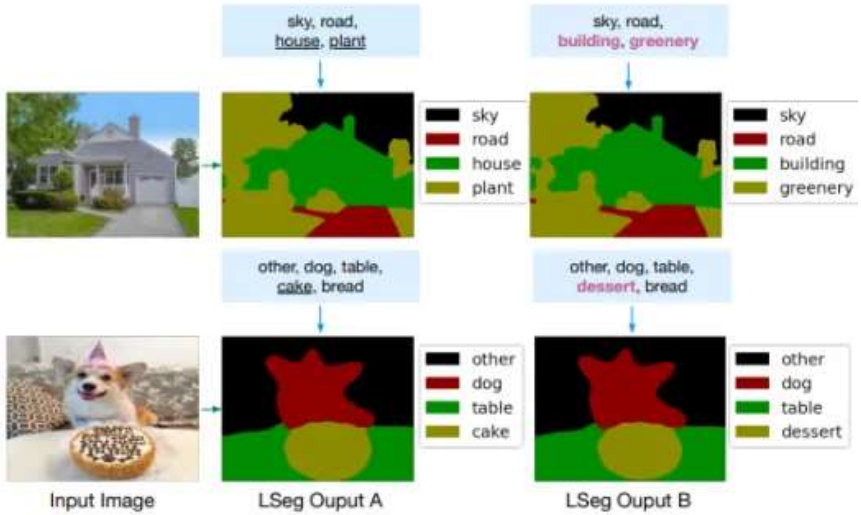
1. 一种是深度可分离卷积(Chollet)的简单级联和非线性激活函数(DepthwiseBlock)
2. 另一种是额外增加了一层针对标签集合的最大池化操作来扩充深度可分离卷积的block(BottleneckBlock)。



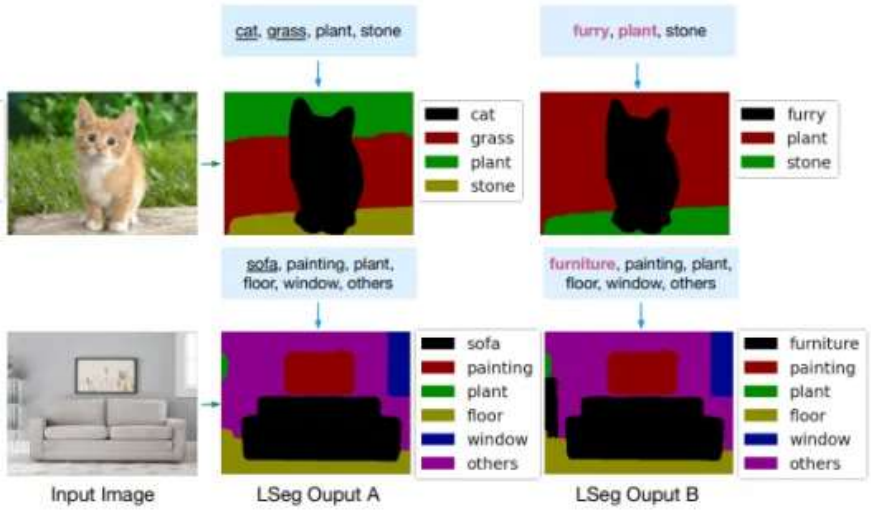

other, dog
other
dog
other, dog, tree
other
dog
tree
other, pet, vehicle, tree
other
pet
tree
other, person, bicycle
other
person
bicycle
other, person, bicycle, sky
other
person
bicycle
sky
other, person, bicycle, sky, mountain, grass
other
person
bicycle
sky
mountain
grass
other, chair
other
chair
other, wall, chair
other
wall
chair
other, wall, chair, window, floor, ceiling
other
wall
chair
window
floor
ceiling
Input Image
(a) LSeg Output
(b) LSeg Output
(c) LSeg Output

案例解说（LSeg识别）

图三右图其中当多个解释与标签集一致时，
模型将重点放在一个最可能的对象上。



(a) Related unseen labels.



(b) Hierarchical unseen labels.

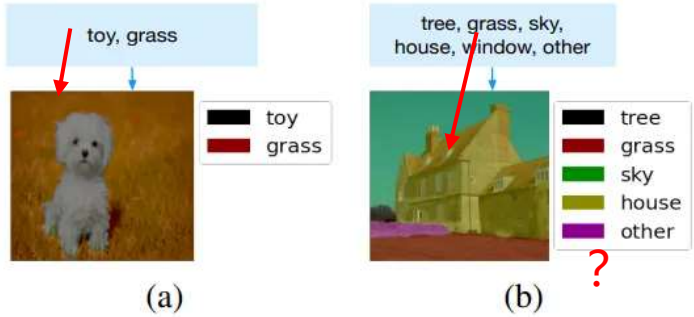


Figure 6: Failure cases.

相关但之前未见过的标签

父级标签

失败案例

图一说明了LSeg泛化到新类的一些突出功能；
该模型在这个不同但语义相关的标签集上产生与之前相似的分割。

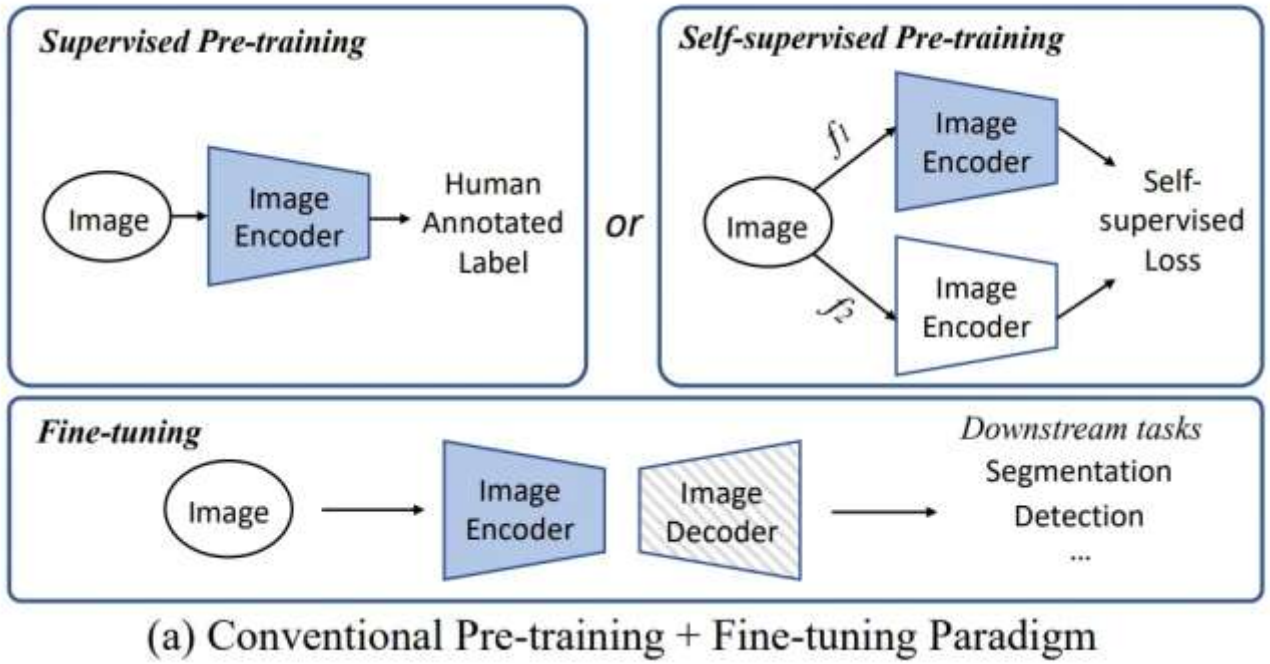
图二表明LSeg可以隐式地为标签层次结构提供正确的分割映射；
标签存在于训练集中，使得模型能够成功地识别“毛茸茸” “家具”这两个父类别。

图三左图说明LSeg只使用来自一个类的正样本进行训练；
当输入标签不包含对应像素的任何真实标签时，模型将最高概率分配给文本嵌入空间中最接近的标签。

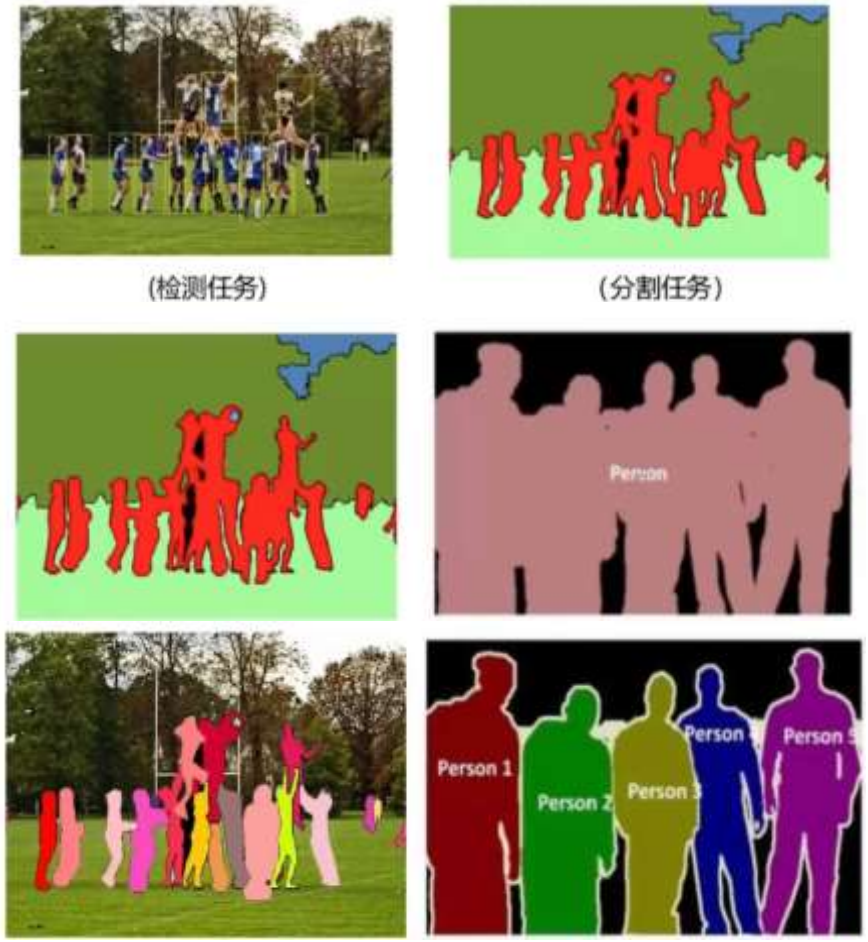
DenseCLIP: Language-Guided Dense Prediction with Context-Aware Prompting

具有上下文感知提示的语言引导的密集预测

研究目标：将CLIP中的**原始图像-文本匹配**问题转换为**像素-文本匹配**问题，并使用像素-文本得分映射来指导密集预测模型的学习。



(a) Conventional Pre-training + Fine-tuning Paradigm



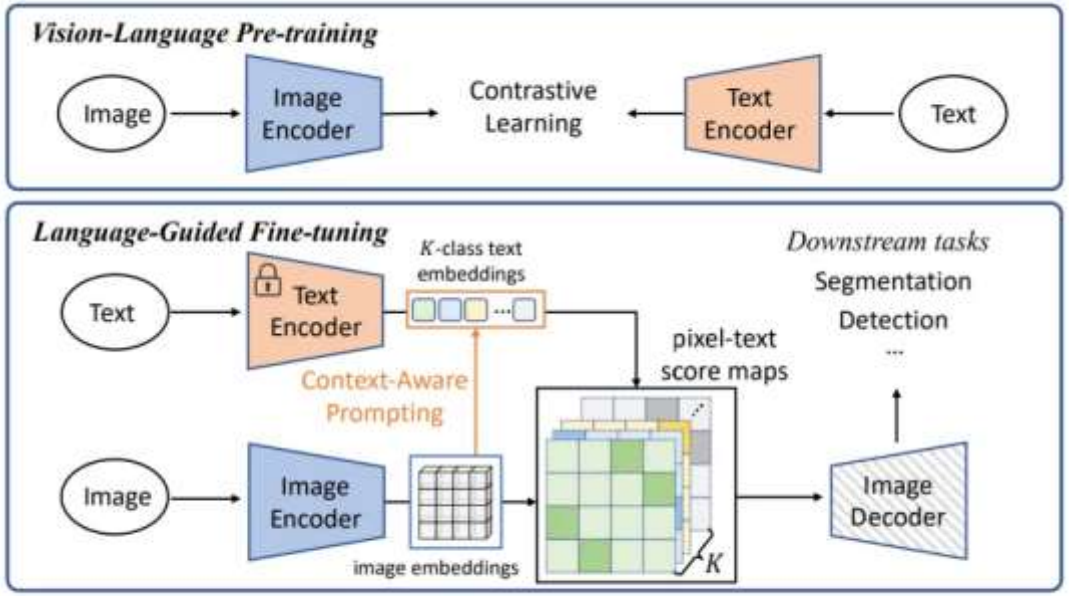
图像分割

语义分割

实例分割

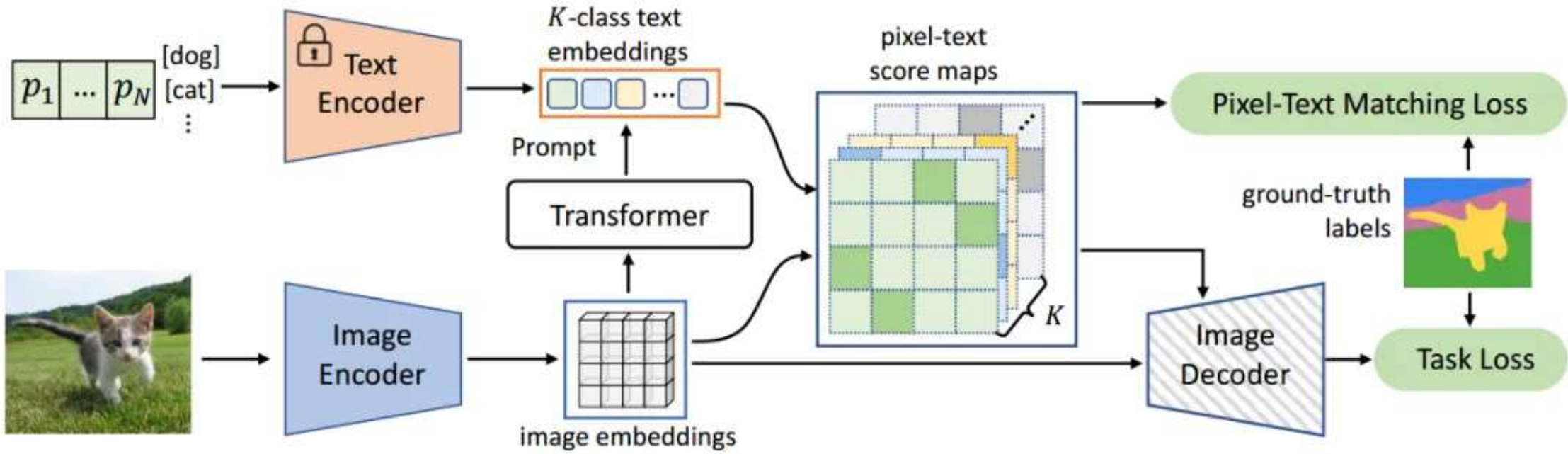
模型总览

图像-标签 预训练 比较学习



(b) DenseCLIP: CLIP Pre-training + Language-Guided Fine-tuning

标签 引导学习 微调



模型原理

语义分割

本文的框架是模型不可知的，可以应用于任何密集的预测pipeline。此外，作者提出使用辅助目标来在分割中更好地利用像素文本得分图。由于得分图 $s \in \mathbb{R}^{H_4 W_4 \times K}$ 可以看作较小的分割结果，因此分割损失为:

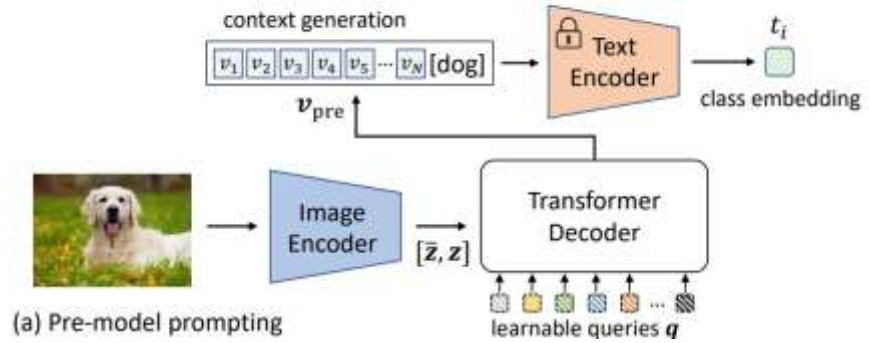
$$\mathcal{L}_{\text{aux}}^{\text{seg}} = \text{CrossEntropy}\left(\text{Softmax}(\mathbf{s}/\tau), \mathbf{y}\right)$$

其中 τ = 0.07是温度系数， $y \in \{1, \ldots, K\}^{H_4 W_4 \times K}$ 是ground truth标签。辅助分割损失有助于恢复特征图的局部性，这有利于分割和检测的密集预测任务。

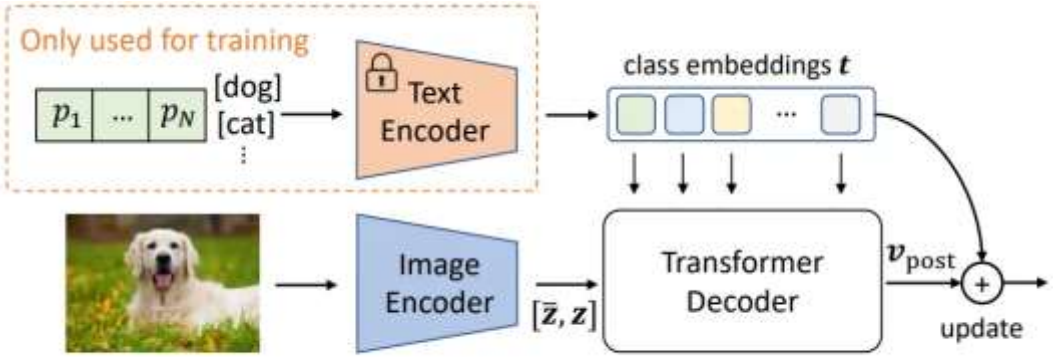
物体检测

在这两种任务下，没有ground truth分割标签。为了构造一个类似于分割中的辅助损失，作者使用边界框和标签来构建一个二元目标 $y \in \{0,1\}^{H_4 W_4 \times K}$ 。辅助目标可定义为二元交叉熵损失:

$$\mathcal{L}_{\text{aux}}^{\text{det}} = \text{BinaryCrossEntropy}(\text{Sigmoid}(\mathbf{s}/\tau), \tilde{\mathbf{y}})$$



(a) Pre-model prompting



(b) Post-model prompting

$$\bar{x}_4 = GAP(x_4)$$

$$[\bar{z}, z] = MHSA([\bar{x}_4, x_4])$$

$$ScoreMap = s = \hat{z}\hat{t}^T, s \in \mathbb{R}^{H_4 W_4 \times K}$$

$$[p, e_k], 1 \le k \le K, p \in \mathbb{R}^{N \times C}, e_k \in \mathbb{R}^C$$

文本编码器之前重新修正文本特征(图像修正文本)

$$v_{pre} = TransDecoder(q, [\bar{z},\ z])$$

q是一组可学习的查询

引入视觉上下文

$t = t + \gamma v_{pre}, \gamma \in \mathbb{R}^C$ 减少文本描述与图像中目标物体的残差

文本编码器之后重新修正文本特征

$$v_{post} = TransDecoder(t, [\bar{z},\ z])$$

$t = t + \gamma v_{post}, \gamma \in \mathbb{R}^C$ 减少文本描述与图像中目标物体的残差

实验结果

后提示参数量小效果好

Pre-train	Language Prompt	V→L Prompt		mIoU (%)	FLOPs (G)	Params (M)
		pre	post			
ImageNet				38.6	227	31.0
CLIP				39.6 (+1.0)	249	31.0
CLIP	✓			42.1 (+3.5)	269	46.5
CLIP	✓	✓		42.9 (+4.3)	368	116.9
CLIP	✓		✓	**43.5** (+4.9)	269	50.2

ChangeCLIP: Remote sensing change detection with multimodal vision-language representation learning

使用多模态视觉语言表示学习的遥感变化检测

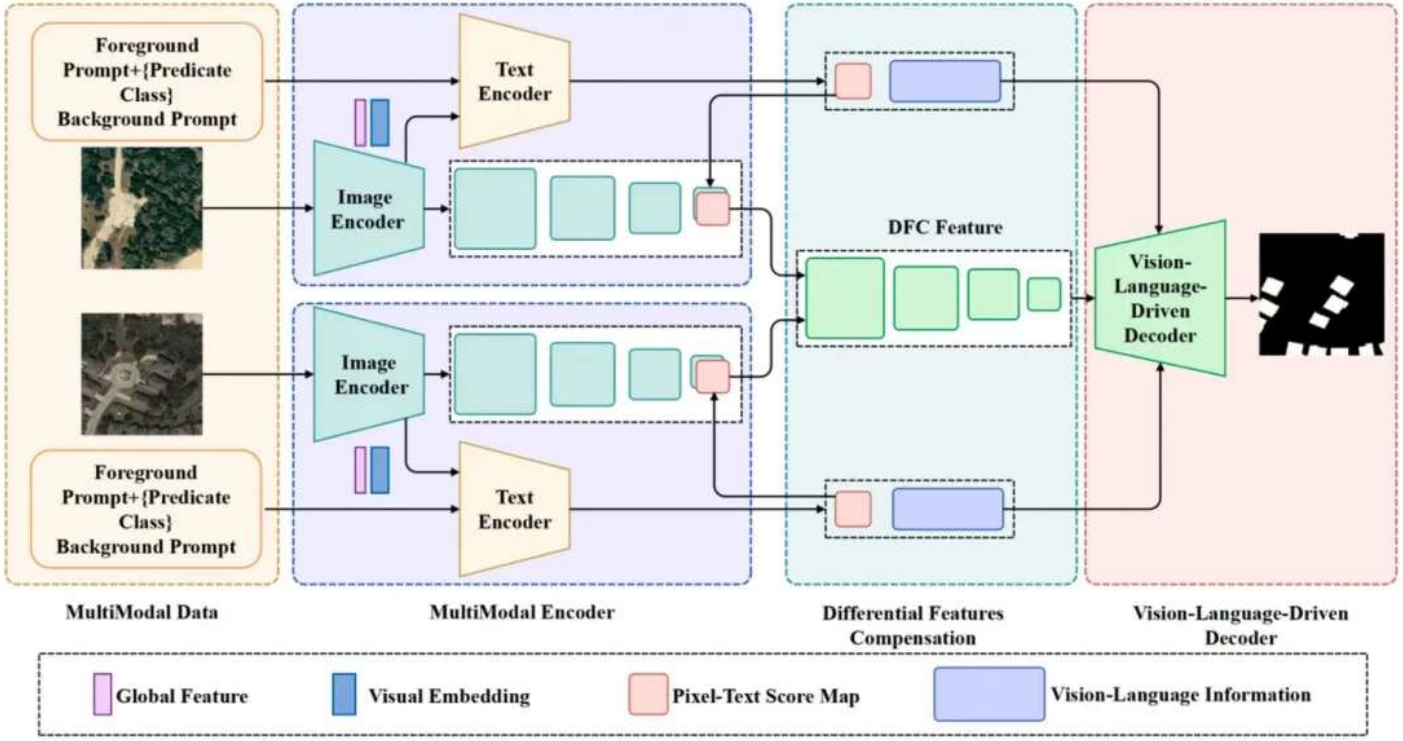
本文提出了一个名为**ChangeCLIP**的**多模态**框架，用于通过整合多模态视觉-语言信息在遥感图像中进行变化检测。

- 提出的**差异特征补偿模块**整合了常用的差异特征计算方法，以优化变化检测中差异特征融合的方式。
- 提出了一种名为**视觉-语言驱动解码器**的多模态变化检测解码方法。它在解码阶段补充语义信息。
- 在解码阶段对文本和视觉特征进行融合使得ChangeCLIP能够生成更准确和全面的表示，提高了变化检测任务的性能。

模型总览

多模态视觉语言表示学习的遥感变化检测

研究目标：现有变化检测方法更注重视觉表示学习，而忽视了多模态数据的潜力。本研究引入了ChangeCLIP，这是一个专门为遥感变化检测（RSCD）定制的新框架，利用图像-文本对中的强大语义信息。



ChangeCLIP模型

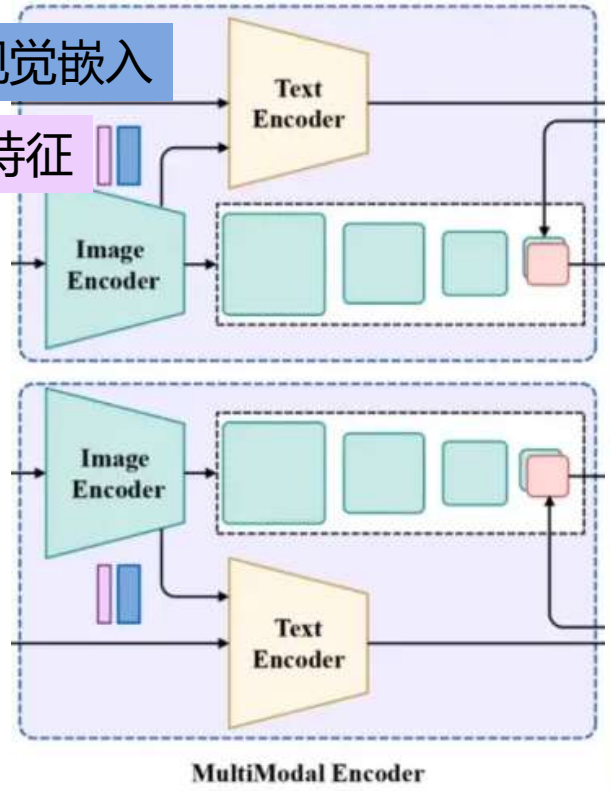
多模态数据：利用CLIP模型的无监督分类能力生成遥感图像的文本提示，从而构建变化检测任务的多模态输入数据。

多模态编码器：我们采用CLIP模型构建**图像和文本编码器**，成为多模态RSCD任务的基础两时间点特征提取器。此外，将图像特征与文本特征整合，有效弥补传统单模态变化检测方法固有的局限性。

差异特征补偿：为了增强模型捕捉两时间点变化的能力，引入了**差异特征补偿（DFC）**模块。该模块采用各种计算方法进行差异特征表示，并采用特征图的加权融合，从而优化对多样的两时间点图像差异的适应性。

视觉-语言驱动解码器：充分利用从编码阶段获得的视觉-语言特征。通过将这些视觉-语言特征与解码阶段的特征结合，设计了一个**以视觉-语言为驱动的解码器。**

多模态编码器



全局特征

$$x_{avg}(i,j) = \frac{1}{HW}\sum_{hw=1}^{HW} x(hw,i,j) = GAP(i,j)$$

视觉嵌入

$$x_1 = Reshape(x)$$

图像编码器

$$x_2 = Concat(x_{avg}, x_1)$$

$$x_3 = x_2 + pos\_embedding$$

$$x_4 = softmax(\frac{x_3 x_3^T}{\sqrt{d_k}})x_3$$ 注意力

$$x_5 = Reshape(x_4)$$

$$global\_feature = x_5[:,:,0]$$

$$visual\_embedding = Reshape(x_5[:,:,1:])$$

$$t_2 = F_{Transformer}(t_1)$$ 文本编码器

$$t_2 = F_{ChangeDetect}(t_2, visual\_embedding)$$ 引入视觉信息

$$t_3 = t_2 + \gamma \times t_2$$

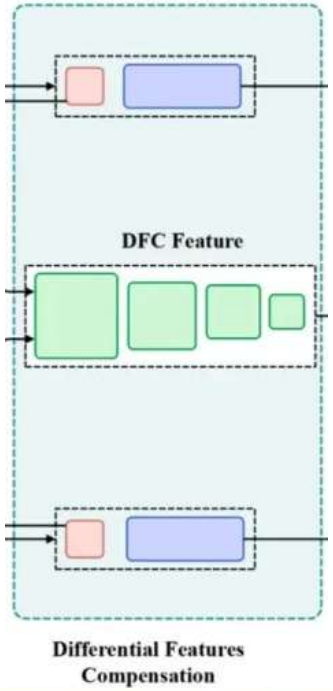
$$score\_map = \eta(t_3) \times \eta(V')$$ 分数预测图

标准化

差异特征补偿(DFC)：Differential features compensation module

首先，**通过减法计算特征图的差异。**为了更好地适应两时间点图像变化的方面，取减法结果的绝对值,并添加一个卷积模块使网络适应优化减法的特征图。因此，得到的特征图突显了变化区域的特征。

同时，考虑到余弦距离可以表示高维空间中不同特征向量的相似性，设计了基于余弦距离的另一个特征差异计算模块。



总体而言，DFC的计算步骤如下：

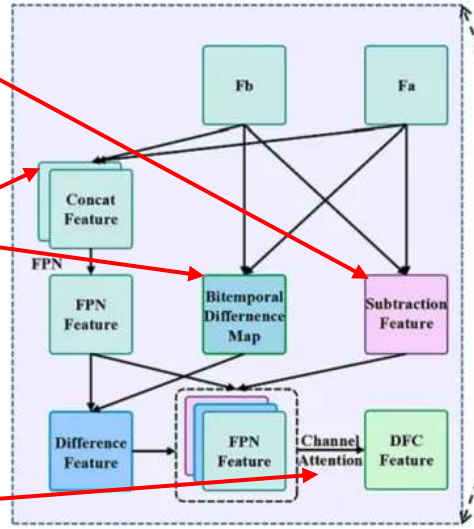
$$F_{sub} = \sigma(abs(F_a - F_b))$$ 作差

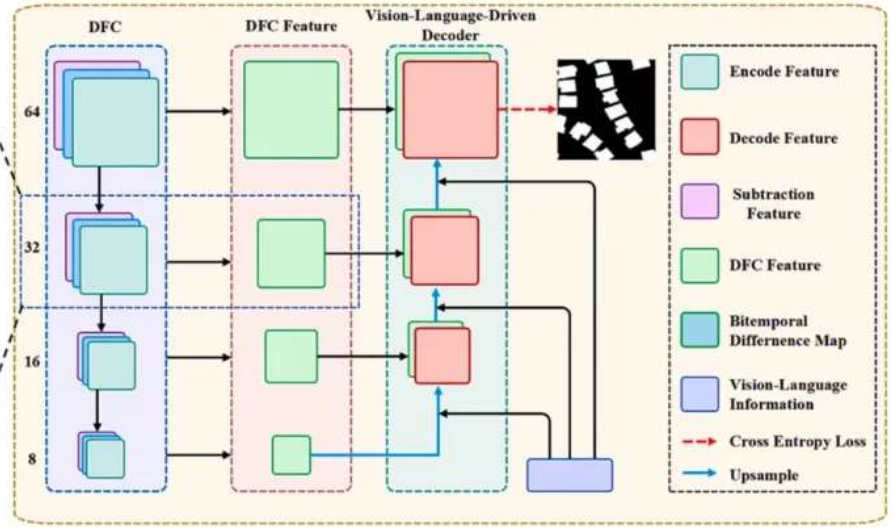
$$\varphi = \frac{x_1 \cdot x_2}{max(||x_1||_2 \cdot ||x_2||_2, \varepsilon)}, x_1 \in F_a, x_2 \in F_b, \in = 1e-8$$

$$m = \phi(1 - \varphi(F_a, F_b))$$ 余弦相似性

$$X = FPN(concat(F_a, F_b))$$

$$X = ChannelAttention(concat(X \times m, F_{sub}, X))$$





视觉语言驱动解码器

低秩双线性注意力模块的计算步骤如下：

$$q = W_I \cdot I \quad k = W_T \cdot T \quad v = W_v \cdot T$$

$$attention\_scores = softmax(\frac{q \cdot k^t}{\sqrt{d_k}})$$

$$output = attention\_scores \cdot v$$

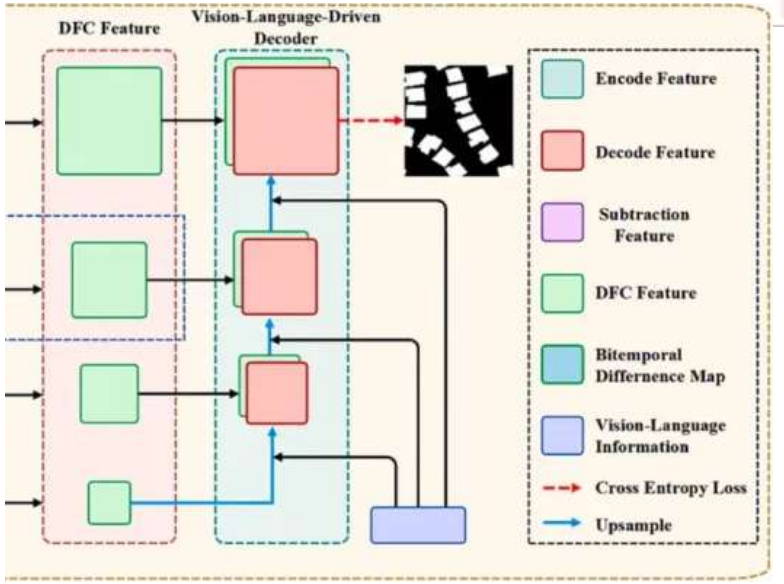
逐层多级特征融合结构其计算步骤如下：

$$F_3 = concat(f(DFC_4, text), DFC_3)$$

$$F_2 = concat(f(F_3, text), DFC_2)$$

$$F_1 = concat(f(F_2, text), DFC_1)$$

$$output = Upsample(F_1)$$



- 利用**Swin Transformer**块在解码阶段建立了全局注意关系，增强了ChangeCLIP的特征表示能力。
- 引入一个低秩双线性注意力模块，提出了一个以视觉-语言为驱动的解码器，将编码阶段的视觉-语言特征与解码阶段的图像特征相结合，补充解码阶段的语义特征。
- 编码阶段的视觉-语言特征学习了来自图像和文本的丰富语义信息。通过将语义信息整合到解码阶段，模型可以更好地识别变化区域并提高整体性能。
- 为了增强**ChangeCLIP**在解码阶段的学习能力，设计了基于**Swin Transformer**块的逐层多级特征融合结构。

实验结果

Table 7
ABLATION STUDY In IoU Index (RN50 / ViT-B/16), Baseline:CLIP Image Encoder Without Text Encoder, ME:Multimodal Encoder, DFC: Differential Features Compensation Module, VLDD: Vision-Language-Driven Decoder, FPN: Feature Pyramid Networks, FCN: Fully Convolutional Network, Sub: Subtraction.

Model	Encode	Neck	Decode Head	LEVIR-CD	LEVIR-CD+	CDD	SYSU-CD	WHUCD
Baseline	Image Encoder	FPN	FCN	75.25/74.73	68.61/68.16	89.74/88.77	61.71/64.82	87.56/79.22
Baseline (Sub)	Image Encoder	FPN (Sub)	FCN	71.19/74.31	60.98/68.75	86.37/83.67	44.13/64.45	80.88/79.73
Baseline + DFC	Image Encoder	FPN + DFC	FCN	75.85/75.24	70.46/68.59	89.86/90.25	69.34/68.42	89.01/82.53
Baseline + ME	Image / Text Encoder	FPN	FCN	76.01/75.35	70.51/69.12	89.91/90.54	65.64/65.64	87.74/82.83
Baseline + ME + VLDD	Image / Text Encoder	FPN	VLDD	84.72/83.41	72.62/71.73	95.72/94.13	68.68/69.76	89.59/82.81
ChangeCLIP	Image / Text Encoder	FPN + DFC	VLDD	85.20/83.99	73.61/75.63	95.87/95.78	70.53/71.41	90.15/90.08