

TS-CAM: Token Semantic Coupled Attention Map for Weakly Supervised Object Localization

Wei Gao, Fang Wan, Xingjia Pan, Zhiliang Peng, Qi Tian

Zhenjun Han, Bolei Zhou, Qixiang Ye

摘要

弱监督目标定位是一个具有挑战性的问题，当给定图像类别标签时，需要学习对象定位模型。优化卷积神经网络 (CNN) 进行分类，倾向于激活局部区分区域，而忽略完整的对象范围，导致部分激活问题。在本文中，我们认为部分激活是由 CNN 的内在特征引起的，其中卷积操作产生局部接受域，并难以捕获像素之间的远程相关性。我们引入了令牌语义耦合注意图 (TS-CAM)，以充分利用 vision transformer 中的自注意机制进行远程相关性提取。TS-CAM 首先将图像分割成一系列补丁标记进行空间嵌入，从而产生远程相关性的注意图，以避免部分激活。然后，TS-CAM 为补丁令牌重新分配与分类相关的语义，使每个标记都能够知道对象类别。TS-CAM 最终将块标记与语义不可知的注意力图耦合，以实现语义感知定位。

关键词：弱监督目标定位；类别激活图；Vision Transformer

1 引言

弱监督学习是指利用带有不完全注释的训练数据学习识别模型的方法。弱监督目标定位只需要图像级别的注释来指示图像中是否存在一类对象，从而学习定位模型。弱监督目标定位吸引了越来越多的关注，因为它可以利用带有标记的丰富 Web 图像来学习对象级模型。

作为弱监督对象定位的基石，类别激活图 (CAM)^[1]利用来自最后一个卷积层的激活映射为对象边界框估计生成语义感知的定位映射。然而，由于分类模型激活的识别区域往往远远小于对象的实际范围^[2]，CAM 对目标区域的估计严重不足。局部判别区域能够最大限度地减少图像分类损失，但难以准确定位目标^[3]。为了解决这一问题，许多人提出了扩散激活^[3-5]或对抗性训练^[4-9]。然而，从根本上解决 CNN 局部表示的固有缺陷的工作却很少。捕获远程相关性 (可以解释为不同空间位置特征之间的语义相关性) 对弱监督目标定位至关重要。

近年来，vision transformer 被引入计算机视觉领域。vision transformer^[10]通过将输入图像分割为带有位置嵌入的小块，并应用级联的变压器块来提取可视化表示，从而构建标记序列。得益于自注意机制和多层感知器结构，vision transformer 可以学习复杂的空间变换，并自适应地反映远程语义关联，这对定位整个对象范围至关重要。然而，由于以下两个原因，视觉变压器不能直接缓解到弱监督目标定位：(1) 在使用块嵌入时，输入图像的空间拓扑被破坏，阻碍了目标定位的激活图的生成。(2)vision transformer 的注意图是语义不可知的 (不能被对象类区分)，不能胜任语义感知的定位。

在本研究中，我们提出了标记语义耦合注意映射 (TS-CAM)，首次尝试使用 vision transformer 进行弱监督目标定位。TSCAM 引入了一个具有两个网络分支的语义耦合结构，如图 1 所示，一个使用块标签执行语义重新分配，另一个在类标签上生成语义不可知的注意力图。语义重新分配，加上类标记

语义激活，使块标记能够感知对象类别。语义不可知注意力图的目的是利用 vision transformer 中级联的自注意模块的优势，捕获块标记之间的远程相关性。TS-CAM 最后将语义感知映射与语义不可知论注意映射进行对象定位。

本工作的贡献如下：

- 提出了标记语义耦合注意映射 (TS-CAM)，作为使用可视化转换器利用远程特征依赖的 WSOL 的第一个坚实基线。
- 提出了语义耦合模块，将语义感知的标记与语义不可知的注意图结合起来，为利用 vision transformer 提取的语义和定位信息进行对象定位提供了一种可行的方法。
- TS-CAM 在两个具有挑战性的弱监督目标定位基准测试中实现了对以前方法的大幅改进，充分利用了 vision transformer 中的远程特征相关性。

2 相关工作

2.1 弱监督目标定位

弱监督目标定位旨在学习仅给定图像级类别标签的对象定位。CAM^[1]是弱监督目标定位的一个代表性研究，它通过使用特定于类的全连接层聚合深度特征图来生成定位图。通过去掉最后一个全连接层，CAM 也可以通过全卷积网络^[11]实现。

大多数方法都是通过在 CAM 中引入复杂的空间正则化技术来扩展激活区域。然而，图像分类与目标定位之间的矛盾一直困扰着他们。根据可视化方法^[12-13]的观察，CNN 倾向于将对象分解为对应于局部接受域的局部语义元素。激活几个语义元素可以带来良好的分类结果。如何从局部接受区域收集全局线索的问题仍然存在。

2.2 弱监督检测和分割

弱监督检测和分割是与弱监督目标定位密切相关的视觉任务。弱监督检测训练网络同时执行图像分类和实例定位^[14-16]。在给定数千个区域建议的情况下，学习过程在训练检测器的同时从包中选择得分较高的实例。以类似的方式，弱监督分割训练分类网络估计伪掩码，这些伪掩码进一步用于训练分割网络。为了生成精确的伪掩码，^[17-20]采用了区域增长策略。同时，一些研究者研究了直接增强特征水平激活区域的方法^[21-22]。另一些则通过多阶段训练^[23]、探索边界约束^[24]、利用等价性进行语义分割^[25]、挖掘跨图像语义^[26]来提炼伪掩模来积累 CAMs。

与 WSOL 类似，许多弱监督检测和分割方法容易对目标部分进行局部定位，而不是整个对象范围。为了系统地解决部分激活问题，需要探索新的分类模型。

2.3 远程相关性

由于 CNN 受限于局部感受野，CNN 不擅长去捕获全局线索。为了缓解这种限制，一种解决方案是利用像素相似性和全局线索来优化激活映射^[17,25,27-28]。另一种解决方法是注意机制^[16]。以自注意的方式将 non-local 操作引入 CNN，以便每个位置的响应是所有 (全局) 位置的特征的加权和。最近的研究在变压器模型中引入了级联自注意机制，以捕获远程特征相关性^[21,29-30]。

3 本文方法

3.1 本文方法概述

本文提出 TS-CAM 方法在训练过的 vision transformer 上生成语义感知的定位映射，如图 1 所示。然而，在 vision transformer 中，只有类标记是语义感知的，而块标记是语义不可知的。为了实现语义感知的定位，本文引入了语义重分配分支，将语义从类标记转移到块标记，并生成语义感知图。将语义感知图与语义不可知注意力图结合，生成语义感知的定位图。

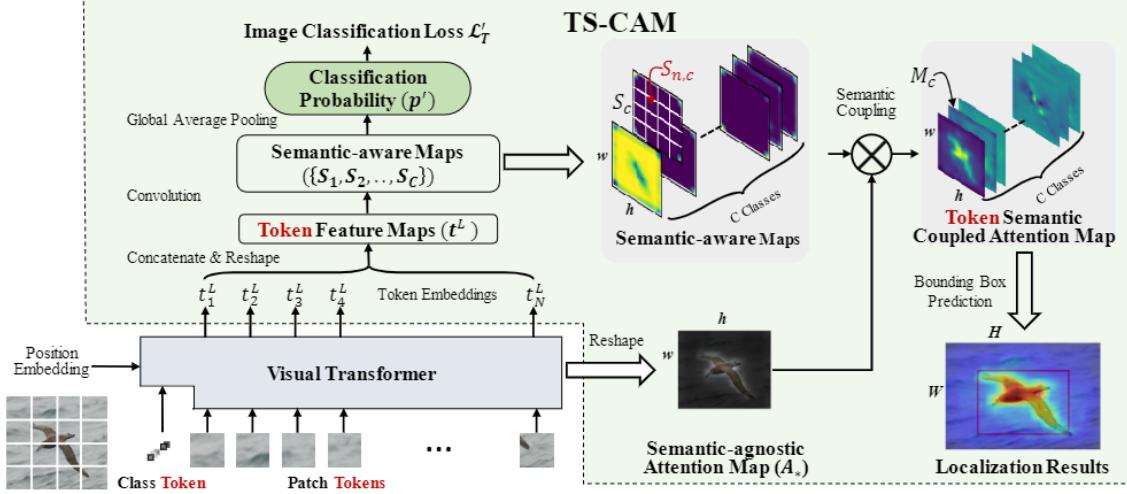


图 1: TS-CAM 框架由特征提取的 vision transformer、语义重分配分支和语义耦合模块组成。注意，沿着语义重分配分支没有梯度反向传播。

3.2 语义重分配

vision transformer 使用类标记来预测图像类别(语义)，同时使用语义不可知的块标记来嵌入对象的空间位置并反映特征的空间依赖性。为了生成语义感知的块标记，我们建议将语义从类标记 t_*^L 重新分配块标记 $\{t_1^L, t_2^L, \dots, t_N^L\}$ 。

如图 1，第 L 个 vision transformer 的快标记嵌入被拼接并转置为 $t^L \in \mathbb{R}^{D \times N}$ 。然后将它们重塑为标记特征映射 $t^L \in \mathbb{R}^{D \times w \times h}$ ，其中 $t_d^L, d \in \{1, 2, \dots, D\}$ 表示第 D 个特征图。类别 c 的语义感知图 S_c 通过卷积计算，如公式 1:

$$S_c = \sum_d t_d^L * k_{c,d} \quad (1)$$

其中 $k \in \mathbb{R}^{C \times D \times 3 \times 3}$ 表示卷积核， $k_{c,d}$ 是一个由 c 和 d 索引的 3×3 卷积核图。 $*$ 是卷积算子。为了产生语义感知图，损失函数定义如公式 2:

$$L_T = -\log \frac{\exp(\sum_n S_{n,y}/N)}{\sum_c \exp(\sum_n S_{n,c}/N)} \quad (2)$$

其中 $S_{n,c}$ 为其是类 c 第 n 个块标记的语义。通过最小化公式 2 将语义重新分配到块标记上，从而生成语义感知图。

3.3 语义不可知的注意力图

为了充分利用视觉转换器的远程特征依赖性，我们提出聚合类标记的注意向量来生成语义不可知的注意力图。 $\mathbf{t}^l \in \mathbb{R}^{(N+1) \times D}$ 指的是第 l 个 transformer 模块的输入，有所有标记的嵌入拼接而成。在第 l

个 transformer 的自注意力操作中，嵌入标记 $\tilde{\mathbf{t}}^l$ 通过下面公式 3 计算得到。

$$\begin{aligned}\tilde{\mathbf{t}}^l &= \text{Softmax} \left((\mathbf{t}^l \theta_q^l) (\mathbf{t}^l \theta_k^l)^\top / \sqrt{D} \right) (\mathbf{t}^l \theta_v^l) \\ &= A_*^l (\mathbf{t}^l \theta_v^l)\end{aligned}\quad (3)$$

其中 $\theta_q^l, \theta_k^l, \theta_v^l$ 分别表示第 l 个 transformer 模块中自注意操作线性变换层的参数。 \top 是卷积操作。 A_*^l 为注意力矩阵， $A_*^l \in \mathbb{R}^{1 \times (N+1)}$ 为该类标记的注意向量。在考虑 K 头的多头注意层中，将公式 3 中的 D 更新为 D' ，其中 $D' = D/K A_*^l$ 则更新为来自 K 个头的注意向量的平均值。

公式 3 表明 A_*^l 通过矩阵乘法运算记录类标记对所有标记的依赖关系。3 表明，自注意操作的类标记的嵌入 $\tilde{\mathbf{t}}^l$ 是通过将其注意向量 A_*^l 与第 l 个 transformer 模块中的嵌入 t^l 相乘来计算的。 $\tilde{\mathbf{t}}^l$ 因此能够看到所有的块标记，其中 A_*^l 表示对每个标记的关注程度。当 2 被优化时，注意向量 A_*^l 被驱动聚焦于对象区域（例如，语义校正的远程特征）进行图像分类。最后的注意向量 A_*^l 的定义如下：

$$A_* = \frac{1}{L} \sum_l A_*^l, \quad (4)$$

它聚合注意向量 A_*^l ，并从级联变压器块中收集特征依赖关系，以表示完整的对象范围。

3.4 语义耦合

由于 A_*^l 是语义不可知的，本文使用逐元素相乘，将 A_*^l 和语义感知图 S_c 耦合，以此得到每个类别的语义耦合注意力图 M_c 。耦合过程如下所示：

$$M_c = \Gamma^{w \times h}(A_*) \otimes S_c \quad (5)$$

其中 \otimes 表示元素的乘法和加法操作。 $\Gamma^{w \times h}(\cdot)$ 表示将注意向量 ($\mathbb{R}^{1 \times N}$) 转换为注意力图 ($\mathbb{R}^{w \times h}$) 的变形函数。 M_c 被上采样到一个语义感知的定位图，该地图使用阈值方法^[31]进行对象包围框预测。

4 复现细节

4.1 与已有开源代码对比

本次复现工作主要引用了 <https://github.com/vasgaowei/TS-CAM> 中的代码。下面是改进部分的关键代码。

```
def norm_cam(cam):
    # cam [B N]
    if len(cam.shape) == 3:
        cam = cam - repeat(rearrange(cam, 'B H W -> B (H W)').min(1,
                                                               keepdim=True)[0], 'B 1 -> B 1 1')
        cam = cam / repeat(rearrange(cam, 'B H W -> B (H W)').max(1,
                                                               keepdim=True)[0], 'B 1 -> B 1 1')
    elif len(cam.shape) == 2:
        cam = cam - cam.min(1, keepdim=True)[0]
        cam = cam / cam.max(1, keepdim=True)[0]
    elif len(cam.shape) == 4:
        B, C, H, W = cam.shape
        cam = rearrange(cam, 'B C H W -> (B C) (H W)')
        cam -= cam.min(1, keepdim=True)[0]
        cam /= cam.max(1, keepdim=True)[0]
        cam = rearrange(cam, '(B C) (H W) -> B C H W', B = B, H=W)
    return cam
```

```

for (image, label) in dataloader:
    logits, x_patch, attn_weights = deit_model(image)

    attn_weights = torch.stack(attn_weights)
    attn_weights = torch.mean(attn_weights, dim=2) # 12 * B * N * N
    feature_map = x_patch.detach().clone() # B * C * 14 * 14
    n, c, h, w = feature_map.shape
    cams = attn_weights.sum(0)[:, 0, 1:].unsqueeze(1).reshape(n, 1, h, w)
    cams = cams.reshape(n, 1, h*w)
    attn_maps = attn_weights.sum(0)[:, 1:, 1:] # B n n
    aggregate_cams = torch.matmul(cams, attn_maps).reshape(n, 1, h, w)
    assign_cams = torch.matmul(attn_maps.unsqueeze(1), cams.unsqueeze(-1)).\
        reshape(n, 1, h, w)
    mix_maps = aggregate_cams + assign_cams
    norm_mix_maps = norm_cam(mix_maps)
    out = norm_mix_maps * feature_map
    return x_logits, out

```

```

for (image, label) in dataloader:
    x_conv, x_trans, attn_weights = conformer_model(image)
    attn_weights = torch.stack(attn_weights)
    attn_weights = torch.mean(attn_weights, dim=2) # 12 * B * N * N

    # conv classification
    x_conv = self.conv_cls_head(x_conv)
    conv_cls = self.pooling(x_conv).flatten(1)

    # trans classification
    x_trans = self.trans_norm(x_trans)
    x_cls = x_trans[:, 0]
    token_cls = self.token_cls_head(x_cls)
    x_logits = conv_cls + token_cls
    feature_conv = x_conv.detach().clone()
    n, c, h, w = feature_conv.shape
    cams = attn_weights.sum(0)[:, 0, 1:].reshape([n, h, w]).unsqueeze(1)
    attn_maps = attn_weights.sum(0)[:, 1:, 1:]
    feature_conv = torch.matmul(attn_maps.unsqueeze(1), feature_conv.reshape(n,
        c, h*w, 1)).reshape(n, c, h, w)
    out = cams * feature_conv # B * C * 14 * 14
    return x_logits, out

```

4.2 实验环境搭建

本文实验的基本环境为 Python3.6 + Pytorch1.7。基于 DeiT^[30]的方法和基于 Conformer^[32]的方法都在两张 Nvidia P100 上运行，批量大小都设置为 128，初始学习率设置为 0.00005，训练 30 轮次后学习率修改为 0.00001，一共训练 60 轮次。

4.3 改进点

4.3.1 改进点一

TS-CAM 中只利用了注意力矩阵中类标记到块标记的注意力向量，但没利用起块标记到块标记的注意力向量，当中蕴含了块标记之间的关系。受到 MCTformer^[33]的启发，他们利用块标记到块标记的注意力向量对类别激活图进行精细化，使得其在弱监督语义分割任务上的性能优于其他现存的方法。本次报告进一步探索如何利用 vision transformer 中自然而然得到的注意力矩阵对类别激活图进行精细

化。基于 TS-CAM，使其在弱监督目标定位任务上有更好优秀的表现。通过分别将语义感知图和语义不可知注意力图可视化，发现语义感知图对整张图片大范围的激活，并没有很好地针对目标区域进行激活。而语义不可知注意力图可以对图像前景进行激活，但激活区域不够完整，同时存在些背景噪声。所以本次报告中通过利用注意力矩阵对语义不可知注意力图进行优化，以此提高定位表现。

在这里，将块标记到块标记的注意力向量称为注意力分配向量，其中的元素表示当前块标记对其他标记的关注程度。将转置后的注意力向量称为注意力聚集向量，其中的元素表示其他块标记对当前块标记的关注程度。分别对语义不可知注意力图进行精细化，得到结果相加后进行归一化得到最终精细化后的语义不可知注意力图。类似的，与语义感知图进行耦合得到定位图。

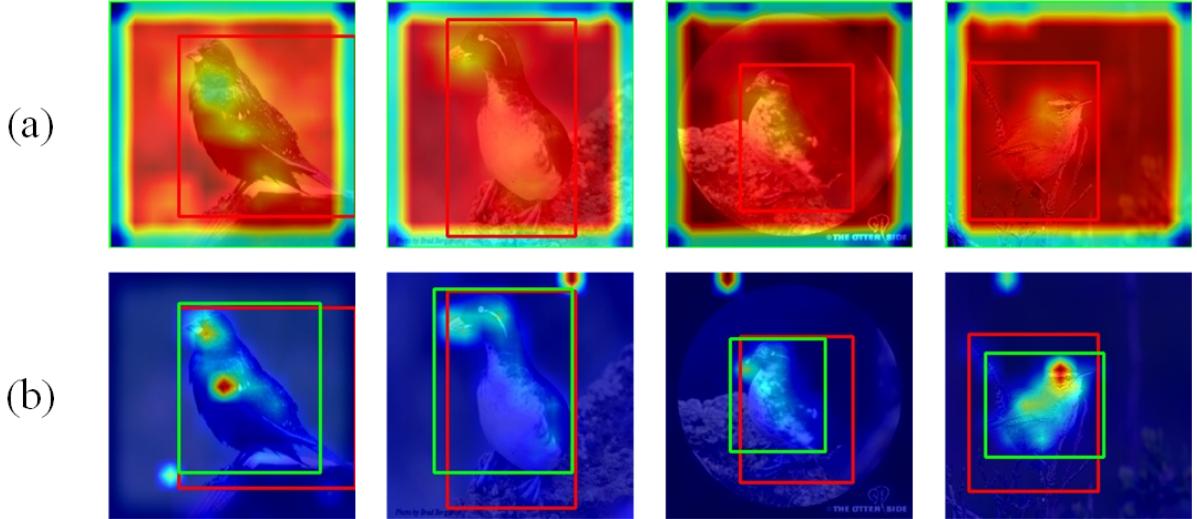


图 2：语义感知图和语义不可知注意力图可视化。其中 (a) 为语义感知图，(b) 为语义不可知注意力图。红色框为标签框，绿色框为预测框。

4.3.2 改进点二

考虑到 ViT, DeiT 等这类 vision transformer 缺乏 CNN 的一些偏置归纳。它们对局部特征的提取能力比不上 CNN，但这种能力对于弱监督目标定位任务来说还是挺重要的。故这里采用 Conformer^[32]作为骨干网络，它由一个 CNN 分支和一个 vision transformer 分支组成，能够同时提取局部和全局特征。

由于 Conformer 有两个分支，本报告打算进一步探索两个分支生成类别激活图的情况。这里对 CNN 分支做出一些调整，最后一个模块不进行下采样，使得 CNN 分支最后的分别率与 vision transformer 分支保持一致。与 vision transformer 分支类似，用一个 3×3 卷积使得最后一个模块的输出的通道数改变为类别数。最后通过一个全局平均层。通过可视化 CNN 分支和 vision transformer 分支的类别激活图，发现 CNN 分支可以正确激活目标区域的同时拥有更少的背景噪声。而 Conformer 的 vision transformer 分支的类别激活图相比于之前 DeiT 的激活效果有改进，但是错误地激活较多背景区域。

基于此观察，决定用 CNN 分支来生成语义感知图。相应地，vision transformer 分支则不再使用块标记来预测分类，而是用回类标签来预测分类。为了使得语义感知图激活得更加完整，这里参考 MCTformer^[33]，用块标记到块标记的注意力向量对语义感知图进行细化，然后再与语义不可知注意力图进行耦合。后面查阅文献发现在弱监督语义分割上有类似的工作^[34]。

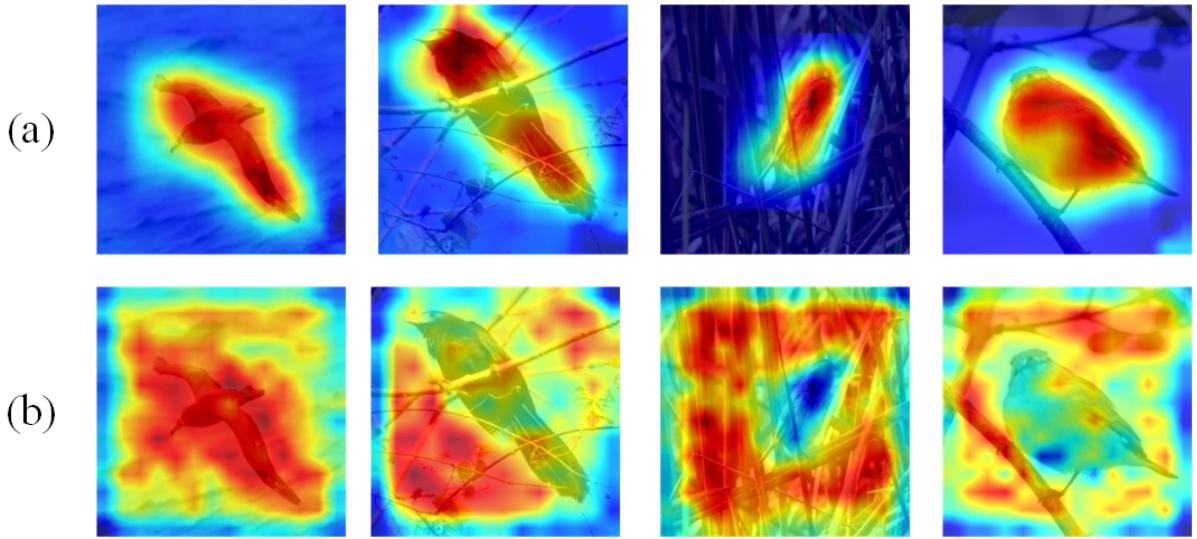


图 3: 两个分支的语义感知图可视化。其中 (a) 为 CNN 分支的结果, (b) 为 vision transformer 分支的结果。

5 实验结果分析

从图 4 得知, 复现结果与论文中报告的性能相近, 后续的改进也有效提升了定位性能。其中 TS-CAM 的 Top-1 定位精度提升 2.7%, Con-CAM 的 Top-1 分类精度提高了 3.4%, Top-1 定位精度提升了 9.1%。各个模型在 CUB-200-2011 数据集上的定位图如图 5 所示。从第二行可以看出, 经过精细化激活的区域相较于之前变得更加完整。第三行展现出双分支骨干网络能力, 结合提取局部特征和全局特征的能力, 使其无需精细化时已经体现出优秀的定位能力。最后一行则体现出转换生成语义感知图的分支和已经对语义感知图精细化的改进效果, 更完整地激活目标区域。

Method	Backbone	Top-1 Cls.ACC	Top-5 Cls.Acc	Top-1 Loc.Acc	Top-5 Loc.Acc	GT-k Loc.Acc
TS-CAM"	DeiT-S	80.3	95.0	71.3	83.8	87.7
TS-CAM*	DeiT-S	79.5	94.5	69.8	82.9	87.0
TS-CAM-refine	DeiT-S	79.5	94.5	73.0	86.0	90.4
TS-CAM"	Conformer-S	81.0	95.8	77.2	90.9	94.1
Con-CAM	Conformer-S	83.7	96.1	80.4	91.5	92.4

图 4: 各个模型在 CUB-200-2011 测试集上的结果。' 表示在 GitHub 中报告的结果; " 表示在论文中报告的结果。* 表示复现的结果。TS-CAM-refine 对应改进点一, Con-CAM 对应改进点二。

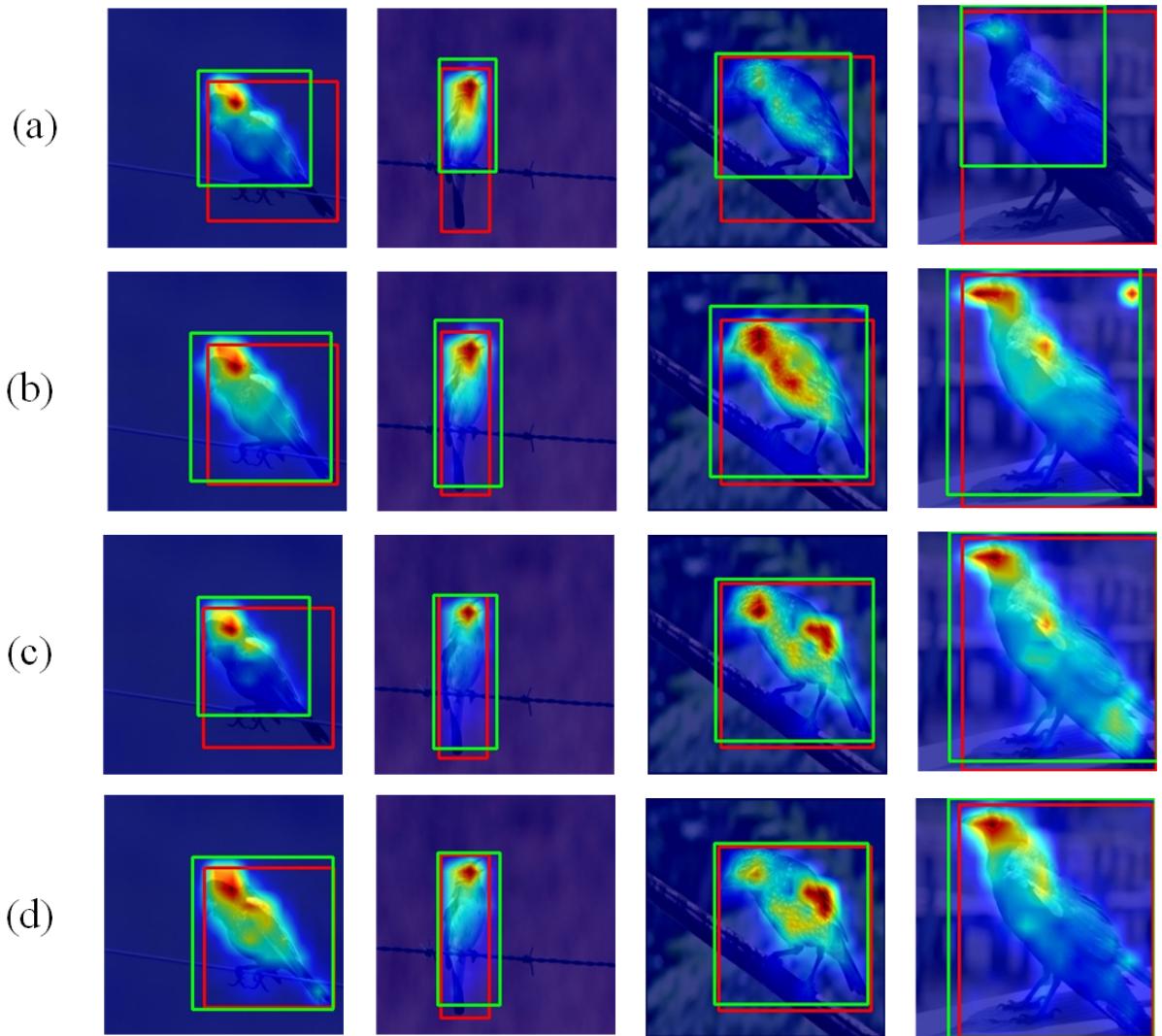


图 5: CUB-200-2011 数据集上定位图可视化。 (a)TS-CAM*; (b) TS-CAM-refine; (c)TS-CAM'; (d)Con-CAM

观察图 6中的表格以及图 7的可视化结果可知，用注意力分配向量来细化语义不可知注意力图的效果很差。虽然这种操作能够扩大激活区域，但同时也提高了对背景区域的激活强度，与语义感知图耦合后，反而使得定位性能更差了。用注意力聚集向量进行细化，可以抑制部分语义不可知注意力图激活强度较弱的背景噪声，以此来获得定位性能的提升，但也可能因此使得激活区域收缩。所以，本次实现中将前面两种策略结合，得到第三种细化策略。通过将前两种的细化策略的结合相加并归一化后，既能使得激活区域更加完整，又能抑制部分背景噪声。

对比图 4以及图 5中的 (c) 与 (d) 可以看出，使用 CNN 分支来生成语义感知图以及用块标记到块标记的注意力向量对其进行细化后，实现了分类与定位性能都有较大幅度的提升，但 GT-Known 稍微下降。

Method	Top-1 Loc.Acc	Top-5 Loc.Acc	GT-k Loc.Acc
TS-CAM*	69.8	82.9	87.0
TS-CAM-AS	15.4	17.7	18.3
TS-CAM-AG	71.7	84.6	89.0
TS-CAM-AM	73.0	86.0	90.4

图 6: 三种精细化语义不可知注意力图的策略。AS 指的是用注意力分配向量进行细化; AG 指的是注意力聚集向量; AM 指的是混合分别用注意力分配向量和注意力聚集向量细化后的结果，并归一化。

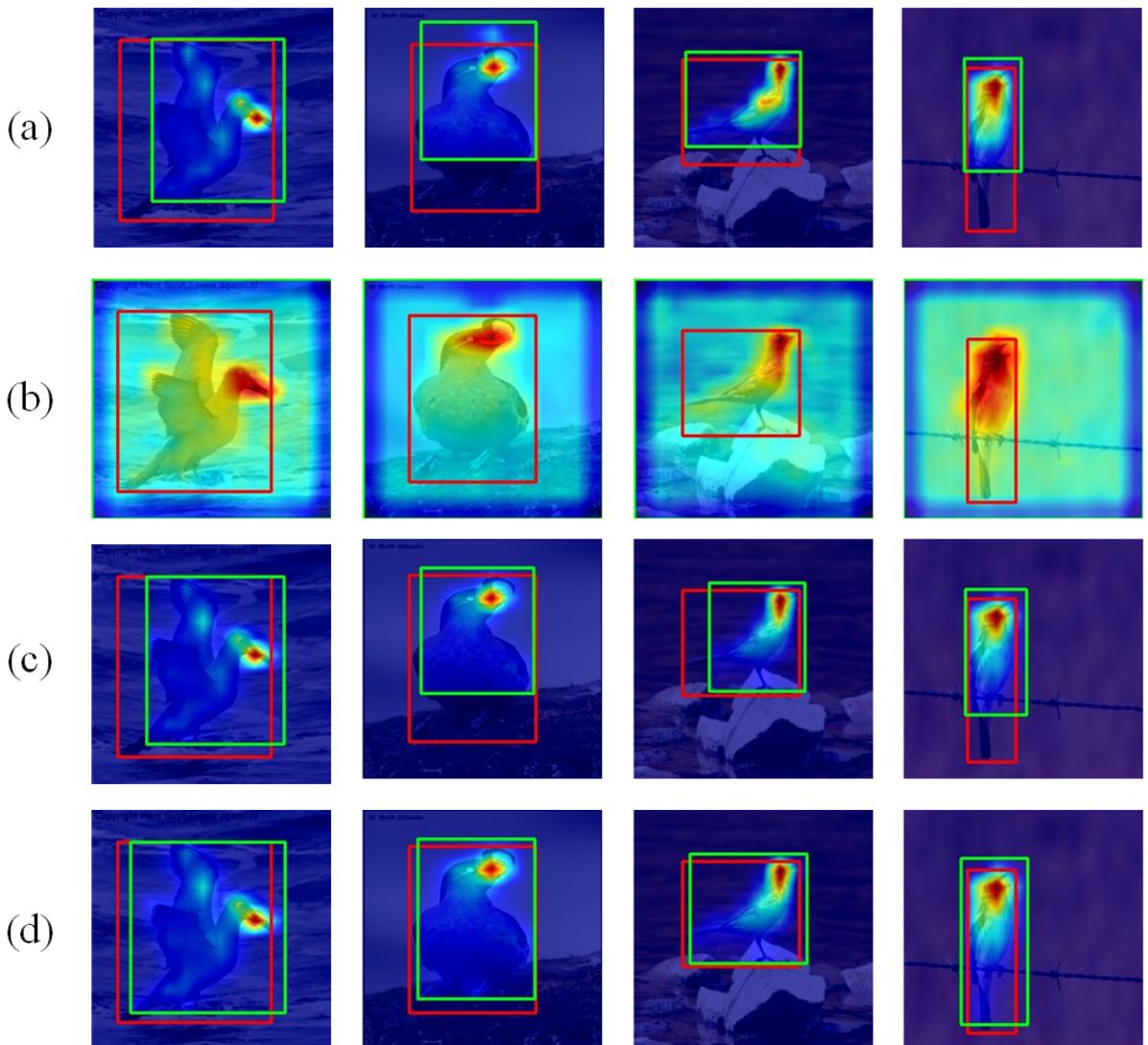


图 7: 可视化用不同策略进行细化的结果。(a)TS-CAM*; (b)TS-CAM-AS; (c)TS-CAM-AG; (d)TS-CAM-AM

图 8展示了改进模型的表现仍然受限的一些情况。观察第一、二列的情况发现，改进模型仍会错误激活水中的倒影。观察第三、四列发现，改进模型仍会对鸟类驻足的地方错误激活。当原始的类别不可知注意力图对非目标区域激活时，对于 TS-CAM-refine 来说，它能够降低对这些区域激活强度，以致于经过阈值处理过后得到更精确的定位图。但如果原始激活强度较大时，效果便不明显了。对于 Con-CAM 来说，可以将这些情况归于受到共现背景的影响。比如说第一、二列的鸟类在数据集中经

常同时出现的背景有水面以及水面中的倒影。对于第三、四列的鸟类在数据集中经常同时出现背景有杆子和树枝。网络在进行分类是不仅是依据不同鸟类的不同特征来进行区分，而且还把共现背景的特征也纳入考虑，最终使得激活图会把一些共现区域也激活了。要解决这种情况，一方面可以加入额外的监督信息，比如文本信息。由于图像级别的标签其监督强度不足，加入额外的监督信息可能有利于解决这种情况。另一方面可以从区分开图像前景与背景的角度入手，比如引入一种抑制背景激活的损失函数来解决。以上提到的不足已经解决思路将作为后续的研究点。

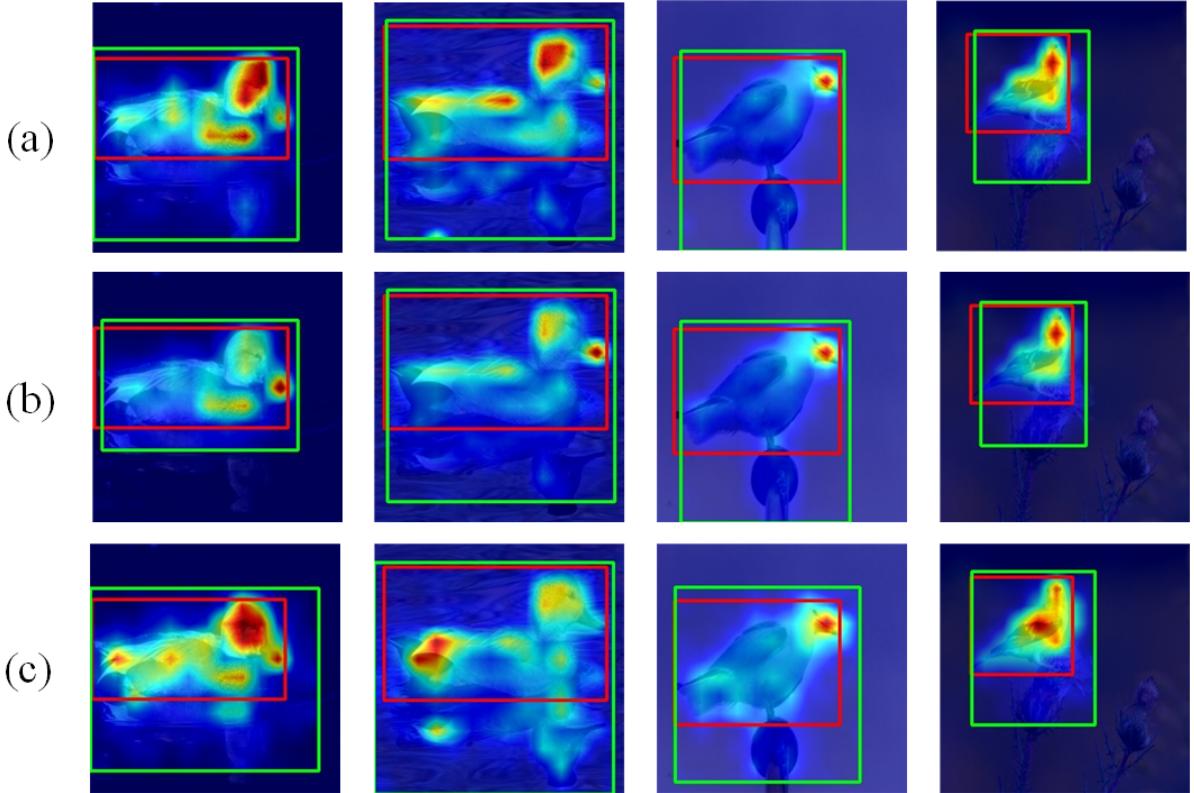


图 8: 可视化改进模型的一些受限情况。 (a)TS-CAM*; (b)TS-CAM-refine; (c)Con-CAM

6 总结与展望

本次报告在复现 TS-CAM 的同时，并基于 TS-CAM 做出一些小改进，在 CUB-200-2011 数据集上取得了更好的定位性能，但当中还存在些不足之处，比如对共现背景的误激活。此外，改进点二在 GT-Known 这一指标上有所下降。如今，多模态领域发展十分迅猛，其中一个名为 CLIP^[35]的多模态模型性能十分优秀。将来打算加入文本监督和抑制背景激活的角度来改进，进一步提升基于 vision transformer 的方法在弱监督目标定位任务上的性能。

参考文献

- [1] ZHOU B, KHOSLA A, LAPEDRIZA A, et al. Learning deep features for discriminative localization[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2921-2929.
- [2] BAE W, NOH J, KIM G. Rethinking class activation mapping for weakly supervised object localization [C]// European Conference on Computer Vision. 2020: 618-634.
- [3] XUE H, LIU C, WAN F, et al. Danet: Divergent activation for weakly supervised object localization[C] // Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 6589-6598.

- [4] KUMAR SINGH K, JAE LEE Y. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 3524-3533.
- [5] YUN S, HAN D, OH S J, et al. Cutmix: Regularization strategy to train strong classifiers with localizable features[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 6023-6032.
- [6] CHOE J, SHIM H. Attention-based dropout layer for weakly supervised object localization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 2219-2228.
- [7] WEI Y, FENG J, LIANG X, et al. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1568-1576.
- [8] ZHANG X, WEI Y, FENG J, et al. Adversarial complementary learning for weakly supervised object localization[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1325-1334.
- [9] MAI J, YANG M, LUO W. Erasing integrated learning: A simple yet effective approach for weakly supervised object localization[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 8766-8775.
- [10] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.
- [11] HWANG S, KIM H E. Self-transfer learning for weakly supervised lesion localization[C]//International conference on medical image computing and computer-assisted intervention. 2016: 239-246.
- [12] BAU D, ZHOU B, KHOSLA A, et al. Network dissection: Quantifying interpretability of deep visual representations[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 6541-6549.
- [13] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks[C]//European conference on computer vision. 2014: 818-833.
- [14] GAO W, WAN F, YUE J, et al. Discrepant multiple instance learning for weakly supervised object detection[J]. Pattern Recognition, 2022, 122: 108233.
- [15] REN Z, YU Z, YANG X, et al. Instance-aware, context-focused, and memory-efficient weakly supervised object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10598-10607.
- [16] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7794-7803.

- [17] WANG X, YOU S, LI X, et al. Weakly-supervised semantic segmentation by iteratively mining common object features[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1354-1362.
- [18] KOLESNIKOV A, LAMPERT C H. Seed, expand and constrain: Three principles for weakly-supervised image segmentation[C]//European conference on computer vision. 2016: 695-711.
- [19] HUANG Z, WANG X, WANG J, et al. Weakly-supervised semantic segmentation network with deep seeded region growing[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7014-7023.
- [20] AHN J, KWAK S. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4981-4990.
- [21] YUAN L, CHEN Y, WANG T, et al. Tokens-to-token vit: Training vision transformers from scratch on imagenet[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 558-567.
- [22] LEE J, KIM E, LEE S, et al. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 5267-5276.
- [23] JIANG P T, HOU Q, CAO Y, et al. Integral object mining via online attention accumulation[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 2070-2079.
- [24] CHEN L, WU W, FU C, et al. Weakly supervised semantic segmentation with boundary exploration[C]//European Conference on Computer Vision. 2020: 347-362.
- [25] WANG Y, ZHANG J, KAN M, et al. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 12275-12284.
- [26] SUN G, WANG W, DAI J, et al. Mining cross-image semantics for weakly supervised semantic segmentation[C]//European conference on computer vision. 2020: 347-365.
- [27] ZHANG X, WEI Y, YANG Y, et al. Rethinking localization map: Towards accurate object perception with self-enhancement maps[J]. arXiv preprint arXiv:2006.05220, 2020.
- [28] ZHANG X, WEI Y, YANG Y. Inter-image communication for weakly supervised localization[C]//European Conference on Computer Vision. 2020: 271-287.
- [29] WU B, XU C, DAI X, et al. Visual transformers: Token-based image representation and processing for computer vision[J]. arXiv preprint arXiv:2006.03677, 2020.
- [30] TOUVRON H, CORD M, DOUZE M, et al. Training data-efficient image transformers & distillation through attention[C]//International Conference on Machine Learning. 2021: 10347-10357.

- [31] ZHANG X, WEI Y, KANG G, et al. Self-produced guidance for weakly-supervised object localization [C]//Proceedings of the European conference on computer vision (ECCV). 2018: 597-613.
- [32] PENG Z, HUANG W, GU S, et al. Conformer: Local features coupling global representations for visual recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 367-376.
- [33] XU L, OUYANG W, BENNAMOUN M, et al. Multi-class Token Transformer for Weakly Supervised Semantic Segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 4310-4319.
- [34] LI R, MAI Z, TRABELSI C, et al. TransCAM: Transformer Attention-based CAM Refinement for Weakly Supervised Semantic Segmentation[J]. arXiv preprint arXiv:2203.07239, 2022.
- [35] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning. 2021: 8748-8763.