LayerCAM: 探索用于定位的分层类激活图

Peng-Tao Jiang* Chang-Bin Zhang* Qibin Hou Ming-Ming Cheng, and Yunchao Wei

摘要

类激活图是从 CNN 的最终卷积层生成的。它们可以突出显示感兴趣类别的判别对象区域。这些发现的对象区域已被广泛用于弱监督任务。然而，由于最终卷积层的空间分辨率较小，这样的类激活图通常只能定位目标对象的粗略区域，限制需要像素级精确对象位置的弱监督任务的性能。因此，我们的目标是从类激活映射以更准确地定位目标对象。在本文中，通过重新思考特征图和它们对应的梯度，我们提出了一种简单而有效的方法，称为 LayerCAM。它可以为 CNN 的不同层生成可靠的类激活图。这个属性使我们能够收集到对象定位信息从粗略（粗略的空间定位）到精细（精确的细粒度细节）级别。我们进一步将它们整合成一个高质量的类激活图，可以更好地突出与对象相关的像素。为了评估由 LayerCAM 生成的类激活图的质量，我们将它们应用于弱监督对象定位和语义分割。实验证明我们的方法生成的类激活图比现有的注意力方法更有效和可靠。源代码可在我们的项目页面获得：https://mmcheng.net/layercam/.

1. 简介

最近，学界已经提出了很多注意力方法 [92, 56, 6] 来利用基于 CNN 的图像分类器来生成类激活图。

这些图能够定位目标物体的区域，其中具有高数值的像素更有可能属于到目标对象。由于图像级标签只告诉目标对象是否存在，它们不提供任何对象位置信息。因此，类激活图的定位能力可以很好地弥补这样一个图像级标签的问题，这进一步促进了图像级监督下的不适定弱监督任务 [33, 1, 21, 20] and weakly-supervised object localization [89, 31] 。

图 1. Grad-CAM [56] 生成的类激活图（顶行）和我们的 LayerCAM（底行）。类激活图是从 conv3_3 和 VGG16 [61] 的 conv5_3 生成的。

类激活图的概念最早在 CAM[92]中提出。他们利用特定的网络结构生成类激活图将图像分类器的全连接层替换为全局平均池化层。后来，Grad-CAM [56] 增强了泛化这种技术的能力，它能够生成类激活任何现成的基于 CNN 的图像分类器的映射。Grad-CAM 利用特征图的平均梯度来表示其对目标类别的重要性。虽然这些方法可以有效定位目标物体，它们之间的一个共同问题是它们都依赖于最终的卷积层来生成类激活图。由于最终卷积层输出的空间分辨率非常低，生成的类激活图只能定位粗略对象区域。如图1所示，Grad-CAM 生成的类激活图从 conv5_3 of VGG-16 [61] 只能定位马大致的位置。他们缺乏获得马的精细细节的能力，例如马腿。然而，弱监督任务，如语义分割，通常需要更准确的对象定位信息。来自最终卷积层的类激活映射仅提供粗略的定位信息弱限制了性能上限监督任务。因此，我们希望获得更细粒度的对象位置来帮助更好地定位目标对象。

由于底部卷积层的输出往往有更大的空间分辨

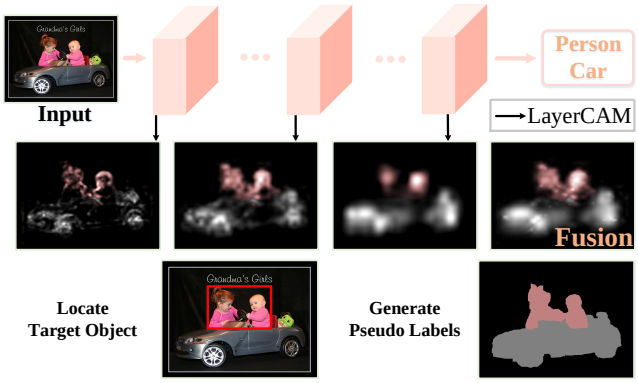


图 2. LayerCAM 图示. LayerCAM 可以应用于任何现成的基于 CNN 的模型并从不同层生成类激活图。不同阶段的类激活图的融合对于对象定位和语义分割任务是有益的。

率，它们对细粒度的细节更敏感。因此，我们尝试将现有的注意力方法应用于它们以生成细粒度的定位信息。Grad-CAM 从 conv3_3 生成的类激活图显示在 图1。可以看出，随着位置的增加，定位变得更糟地图中的强值散布在整个图像周围。我们分析 Grad-CAM 只考虑捕捉全局信息每个特征图，导致其中的局部差异丢失。我们在 第3.1节中提供了更多讨论来分析这个问题为什么 Grad-CAM 在底部卷积层中变得更糟。

为浅层生成可靠的类激活图获得更准确的细粒度目标定位信息，我们在本文中提出了一种简单而有效的方法 LayerCAM。具体来说，我们重新思考了特征图之间的关系以及它们对应的梯度。与之前的注意力方法只考虑每个特征图的全局信息不同，我们利用梯度来突出显示特征图中每个位置对感兴趣的类别的不同重要性。通过这样的操作，可以有效保留目标物体的细粒度细节，同时去除背景中的细节。通常，LayerCAM 具有以下优点：

- LayerCAM 不仅可以从最终的卷积层生成可靠的类激活图，还可以从浅层生成可靠的类激活图，在那里我们可以获得粗略的空间位置和细粒度的对象细节。
- 来自不同层的类激活图通常是互补的。这一优势促使我们将它们结合起来以生成更精确和完整的特定于类的对象区域，这将显着有益于弱监督任务。
- LayerCAM 易于应用于现成的基于 CNN 的图像分类器，无需修改网络架构和反向传播方式，使其更通用且使用方便。

为了证明类激活图的质量，我们将它们应用于弱监督对象定位和语义分割任务。我们的 LayerCAM 的插图显示在 图2。在这两个任务上的实验表明，我们的方法比以前的注意力方法实现了更好的对象定位能力，证明了我们的 LayerCAM 的有效性。此外，从基于 CNN 模型的浅层生成细粒度目标位置的特性也可用于准确定位工业图像中的微小缺陷。

2. 相关工作

2.1. 注意力方法

研究人员提出了许多注意力方法 [60, 63, 85, 93, 50] 来从强大的基于 CNN 的图像分类器中定位感兴趣类别的对象区域 [19, 16, 79, 84, 24, 66, 39]. 注意力方法 [92, 56] 或注意力模组 [9, 71, 72]的有效性使一众视觉任务收益。[83, 58, 8, 46, 22, 45, 10] 在这里，我们主要讨论与我们的工作高度相关的两种注意力方法。

类激活映射。 这类注意力方法 [92, 56, 6, 70] 从最终的卷积层生成类激活图。我们在 图3 中展示了这些方法的一般过程。类激活图是通过将每个特征图乘以其权重，然后对所有加权特征图进行求和来获得的。最后，应用 ReLU 操作来过滤掉负激活。

这些注意力方法之间的区别在于为每个特征图生成权重的方式。CAM [92] 从全连接中获得权重层。他们用全局平均池化层替换了图像分类器中的第一个全连接层。Grad-CAM [56] 将特定于类的梯度流到每个特征图，然后将每个特征图的梯度平均作为其权重。在 Grad-CAM++ [6]中，与 [56]相似，他们还利用特征图的梯度来生成其权重。Score-CAM [70] 摆脱对梯度的依赖，并通过其转发分数为每个特征图生成权重。 上述方法的一个共同点是它们都从最终的卷积层生成可靠的类激活图。与这些方法不同，我们的 LayerCAM 可以从 CNN 的不同层生成可靠的类激活图。

赢家通吃方法。 [86] 提出了一种自上而下的反向传

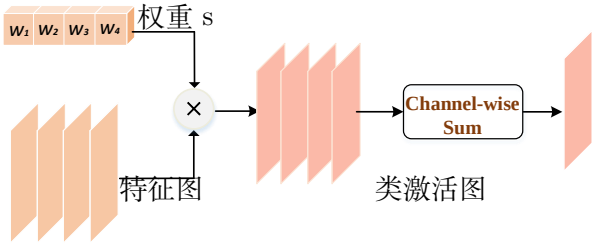


图 3. 类激活映射方法的过程 [92, 56, 6]。

播方案，c-MWP，它基于赢家通吃方法的概率向下传递网络中的信号。它可以为图像分类器的所有卷积层生成类激活图。然而，正如 [56] 中所验证的那样，来自最后一个卷积层的 c-MWP 映射不如 Grad-CAM 忠实，后者很少用于弱监督任务。而且，这种自顶向下的过程很复杂，同时比 Grad-CAM 需要更多的运行时间。最近，PRM [94] 利用局部最大值，即峰值，作为顶部信号，以赢家通吃的方式向下反向传播网络，以提取详细的类实例激活图。

尽管它们能够生成分层类激活图，但从这些图中提取的线索通常涵盖小对象区域。当使用赢家通吃方法时，只保留最相关的神经元。此外，NormGrad [52] 利用相同层为不同层生成类激活图。然而，NormGrad 更有可能捕获小的和有辨别力的对象区域，而不是完整的对象区域。与它们不同的是，LayerCAM 生成的类激活图往往比它们覆盖更多的对象区域。此外，LayerCAM 易于应用于现成的基于 CNN 的图像分类器，而无需修改网络架构，使得类激活图更容易获得。

2.2. 分层语义

许多视觉任务，例如具有挑战性的物体检测任务 [38, 41]，显著性物体检测任务 [40, 73]，和语义分割任务 [91, 36], 都受益于不同特征层次的语义知识。此外，谢等人 [80] 通过利用来自不同层次的 CNN 的特征，在很大程度上改进了边缘检测任务。王等人 [74, 75] 用层次结构对人类解析进行建模。对于视觉对象跟踪，沈等人 [57] 充分利用了不同层次的特征并将它们融合以获得更好的跟踪结果。我们的 LayerCAM 还利用了来自不同层的分层语义知识。从浅层生成的类激活图倾向于捕获目标对象的细粒度细节。而从深层生成的类激活图通常定位粗略的空间对象区域。来自不同层次结构的类激活图都有助于定位目标对象。

2.3. 弱监督对象定位

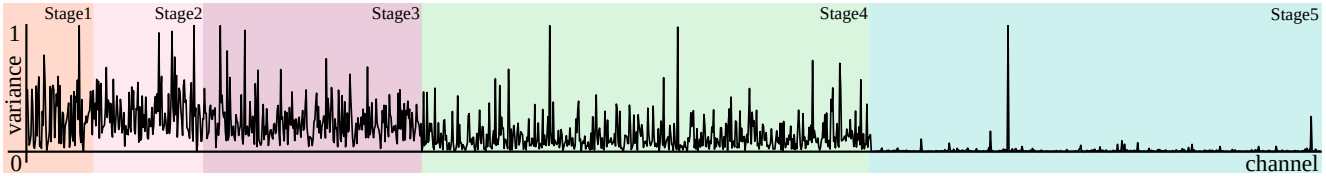
弱监督对象定位 (WSOL) 仅使用图像级标签来查找目标对象的紧密框。一些研究人员 [32, 11, 17, 15, 29, 69] 试图将 WSOL 作为多实例学习框架。另一种方法 [95, 27, 68, 81] 从对象提议先验中选择合适的对象紧框 [51, 42]。

最近很多使用注意力的 WSOL 方法 [?, 92, 89, 90, 30, 82, 62]已经被提出，例如 CAM[92]、Grad-CAM[56]和 ACoL [89]。

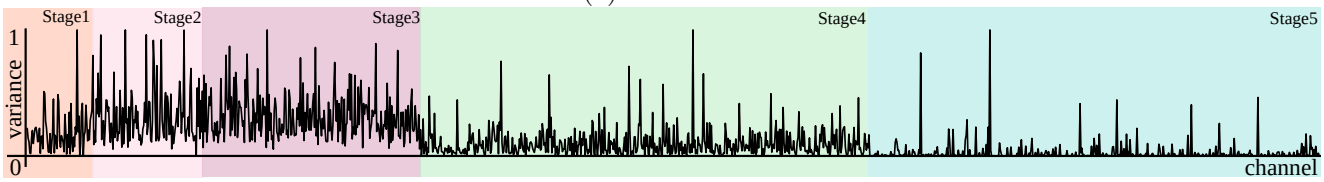
CAM 和 Grad-CAM 通过提取类激活图中的置信区域来识别对象区域。然而，定位性能是有限的，因为置信区域通常很小而且很粗糙。金等人，[30] 利用两个训练步骤来查找不同的对象定位信息。张等人，[89] 使用两个 CNN 分类分支基于擦除策略从类激活图中找到更可信的区域。基于注意力方法的定位方法都从 CNN 的最终卷积层生成类激活图。与以前的定位方法不同，我们尝试通过为不同的卷积层生成可靠的类激活图并寻找更细粒度的定位信息来帮助准确定位目标对象，从而挖掘更完整和准确的对象位置。

2.4. 弱监督语义分割

具有图像级标签的弱监督语义分割 (WSSS) 已被广泛研究，因为图像级标签无需太多人工即可轻松获得。由于图像标签仅提供某个类别的存在，而没有任何空间位置信息，因此该任务仍然是一个具有挑战性的问题。尽管困难重重，但过去几年已经提出了很多 WSSS 工作 [87, 59, 34, 53, 88, 28, 44]。一些工作 [47, 49, 2] 利用图像标签直接训练分割模型。此外，由于注意力方法的流行 [92, 56, 6], 许多 WSSS 方法 [76, 54, 21, 77, 25, 26, 65, 14, 35] 使用从类激活图中提取的对象定位线索。他们首先生成伪分割标签，然后使用它们来训练分割模型。定位对象的完整性在很大程度上影响伪分割标签的质量。



(a) GradCAM



(b) GradCAM++

Image Conv1_2 Conv2_2 Conv3_3 Conv4_3 Conv5_3

(c) 特征图

图 4. (a-b) 展示了 VGG16 不同阶段每个特征图对应的梯度的方差。(c) 展示了从不同阶段随机选择的特征图。

我们从 CNN 的不同层融合的类激活图可以发现更多完整和准确的对象区域，有利于 WSSS 任务。

2.5. 表面缺陷定位

利用计算机辅助工具检查工业产品的质量是提高工业生产质量和效率的重要途径。为了找到工业图像中的表面缺陷位置，许多研究人员 [12, 64, 67, 43] 经常使用完全监督的方法。具体来说，他们需要在工业图像中标注缺陷的位置，然后训练一个分割或检测网络。虽然这类方法取得了非凡的性能，但标记缺陷却相当困难。因为表面上的缺陷及其周围的补丁往往具有非常低的对比度，例如 图7(a)，在工业图像中标记缺陷变得具有挑战性。

此外，在特定场景下检查工业图像中的失败往往需要专业知识，需要大量的人力和时间。因此，弱监督方法值得研究，因为它们可以显着降低注释成本。我们利用从浅层生成的类激活图来定位工业图像中各种形状的微小缺陷，因为来自浅层的图对细粒度的对象细节很敏感。

3. **方法**

在本节中，我们首先回顾两种最相关的方法，即 Grad-CAM 和 GradCAM++。然后我们介绍我们的方法，LayerCAM。

3.1. 回顾 **Grad-CAM** 和 **Grad-CAM++**

形式上，让 f 表示图像分类器，θ 表示其参数。对于给定图像 I，将其输入分类器，我们可以通过下式获得目标类别 c 的预测分数 y^c：

$$y^c = f^c(I, \theta). \tag{1}$$

A 为 CNN 中最终卷积层的输出特征图，A_k 为 A 中的第 k 个特征图。预测得分 y^c 相对于特征图 A_k 中空间位置 (i, j) 的梯度可以通过 $g_{ij}^{kc} = \frac{\partial y^c}{\partial A_{ij}^k}$ 获得。要生成目标类别 c 的类激活图，Grad-CAM 和 Grad-CAM++ 为每个特征图 A_k 分配一个 channel-wise 权重 w_k。然后对特征 A 中的所有特征图进行线性加权求和。最后，应用 ReLU 操作从类激活图中去

除负面响应，公式为：

$$M^c = \text{ReLU}\left(\sum_k w_k^c \cdot A_k\right). \tag{2}$$

Grad-CAM 通过对特征图 A_k 中所有位置的梯度求平均值来获得通道级别参数 w_k^c，表示为：

$$w_k^c = \frac{1}{N}\sum_i \sum_j g_{ij}^{kc}, \tag{3}$$

其中 N 代表特征图 A_k 中地点的数量。对于 Grad-CAM++ [6]，通道级别参数 w_k^c 可以通过计算下式获得：

$$w_k^c = \sum_i \sum_j \alpha_{ij}^{kc} \cdot relu(g_{ij}^{kc}), \tag{4}$$

其中 α_{ij}^{kc} 由下式得到：

$$\alpha_{ij}^{kc} = \frac{\left(g_{ij}^{kc}\right)^2}{2\left(g_{ij}^{kc}\right)^2 + \sum_a \sum_b A_{ab}^k (g_{ij}^{kc})^3}, \tag{5}$$

其中 (a, b) 代表 A_k 中的空间地点。Grad-CAM 和 Grad-CAM++ 的区别在于后者利用特征图和梯度来生成通道级别参数。Grad-CAM++ 在出现多个对象实例时表现出更好的对象定位能力。

虽然 Grad-CAM 和 Grad-CAM++ 可以从最终的卷积层生成可靠的类激活图，但定位的对象区域通常很小而且很粗糙。

我们希望找到更多细粒度的定位信息来修复来自最终卷积层的类激活图，从而更好地定位目标对象。众所周知，CNN 的浅层具有更大的空间分辨率，这使得它们能够捕获目标对象的更细粒度的细节。因此，获得细粒度对象细节的自然想法是将 Grad-CAM 或 Grad-CAM++ 应用于浅层。然而，根据我们的实验，由 Grad-CAM 或 Grad-CAM++ 生成的来自浅层的类激活图通常包含许多误报，如 图1 所示。下面，我们首先分析为什么 Grad-CAM 和 Grad-CAM++ 无法为浅层生成可靠的类激活图，然后介绍我们的方法 LayerCAM。

分析 Grad-CAM 和 Grad-CAM++ 都为第 k 个特征图 A_k^c 分配全局权重 w_k^c，其中 A_k^c 中的每个位置具有相同的权重 w_k^c。然而，浅层中的特征图倾向于捕获细粒度的细节，无论它们属于目标对象还是背景，如 图4(c) 所示。因此，全局权重不能消除背景中的噪声区域，这使得生成的类激活图无法准确定位目标对象。

此外，我们还对全局权重是否可以代表一个特征图中每个位置的重要性进行了数值分析。对于 Grad-CAM，我们计算梯度 g^{kc} 的方差，其中方差表示每个位置的梯度与平均梯度 w_k^c 的差异。对于 Grad-CAM++，我们计算 $\alpha^{kc} \cdot relu(g^{kc})$ 的第 k 个特征图的方差。我们从 VGG16 中选择每个阶段的最后一个卷积层。如图4(a-b) 所示，在最后阶段，我们可以看到大多数特征图对应的方差趋于零。这表明特征图中每个空间位置的权重大约等于全局权重。因此，在最后阶段，Grad-CAM 和 Grad-CAM++ 使用的全局权重都可以表示特征图中每个空间位置的重要性。然而，在浅层，大多数特征图对应的方差非常大。全局权重不能代表特征图中不同位置在目标类别上的重要性。因此，Grad-CAM 和 Grad-CAM++ 无法为浅层生成可靠的类激活图。

3.2. LayerCAM

基于上述分析，我们提出了 LayerCAM，它能够以非常简单有效的方式为所有层收集可靠的类激活图。具体来说，为了为特征图中的每个空间位置生成单独的权重，我们利用后向特定于类的梯度。

正如 [6] 中的经验验证，与特征图中某个位置对应的正梯度表明增加该位置的强度将对目标类别的预测分数产生积极影响。对于具有正梯度的位置，我们使用它们的梯度作为权重。那些具有负梯度的位置分配为零。

形式上，第 k 个特征图中空间位置 (i, j) 的权重可以写为：

$$w_{ij}^{kc} = relu(g_{ij}^{kc}). \tag{6}$$

为了获得某一层的类激活图，LayerCAM 首先将特征图中每个位置的激活值乘以一个权重：

$$\hat{A}_{ij}^k = w_{ij}^{kc} \cdot A_{ij}^k. \tag{7}$$

最后将结果 $\hat{A}_k$ 沿通道维度线性组合，得到类激活

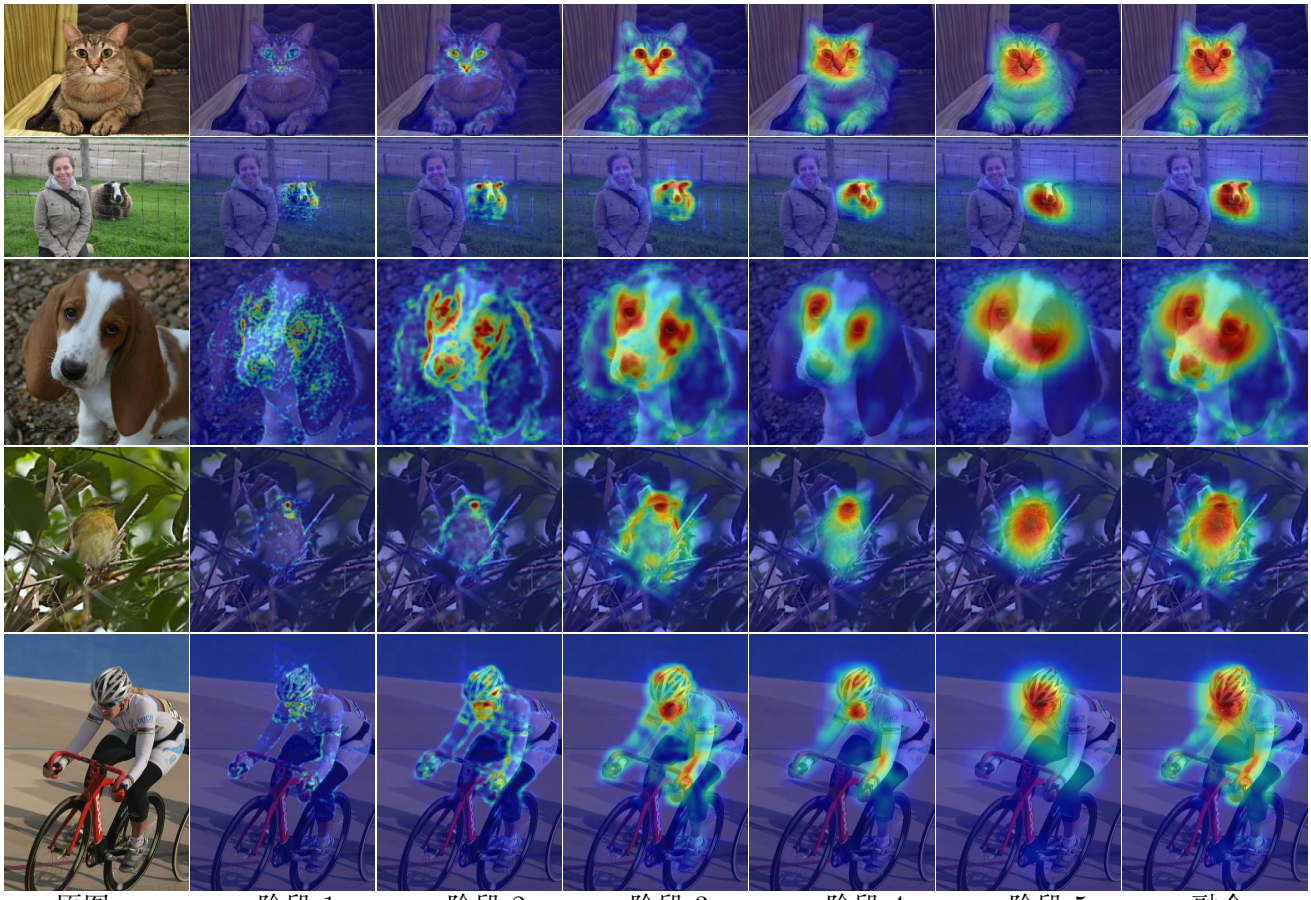


图 5. 不同阶段的类激活图的比较。图像是从 PASCAL VOC 数据集中随机选择的 [13]。**阶段** 5 代表类激活图是从 VGG16 中第 5 阶段的最后一个卷积层生成的。**融合**表示从阶段 3、阶段 4 和阶段 5 融合的类激活图。注意，阶段 1、阶段 2 和阶段 3 的类激活映射根据 Eqn. (9) 进行缩放。

图，公式如下：

$$M^c = \mathrm{ReLU}\left(\sum_k \hat{A}^k\right). \tag{8}$$

基于上述操作，从浅层生成的类激活图可以捕获可靠的细粒度对象定位信息，如图5所示。我们认为这主要得益于不仅考虑了不同渠道的重要性，而且还考虑了不同空间位置的重要性。每个位置的单独权重可以反映不同位置在与目标类别相关的特征图中的重要性。我们将在实验部分进行更多的定性和定量分析。

4. 实验

本节首先进行弱监督目标定位实验，验证 LayerCAM 的定位能力。然后我们利用图像遮挡实验来测试来自最终卷积层的类激活图的一般定位能力的可靠性。此外，我们进行了表面缺陷检测实验，以表明来自浅层的类激活图可以找到细粒度的对象定位信息。最后，我们证明了来自不同阶段的类激活图的组合有利于弱监督语义分割。

4.1. 弱监督目标定位

ILSVRC 基准 [55] 中提出了对象定位实验，旨在为最高预测类别定位对象边界框。我们在具有 50000 张图像的 ILSVRC 验证集上评估我们的方法的定位能力。定位精度由 loc1 和 loc5 指标衡量。loc1 度量表示如果估计的边界框和真实边界框之间的交集（IoU）大于或等于 0.5，同时 top1 预测类是正确的，则估计结果属于正确的类别类是正确的。loc5 指标用于前 5 个预测类别。

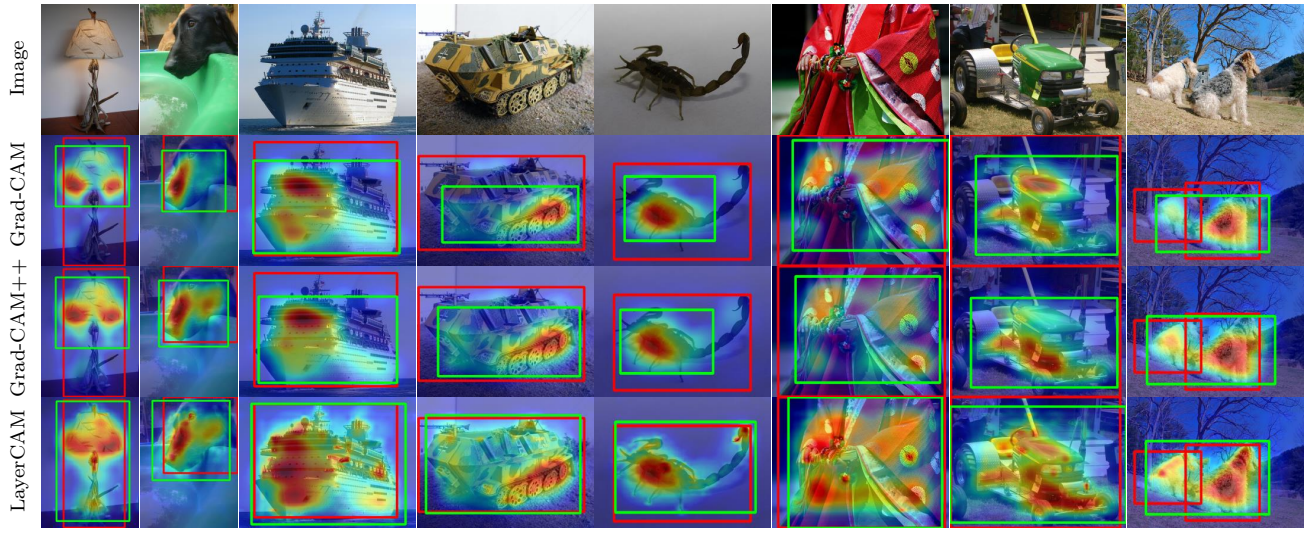


图 6. 不同方法之间定位结果的比较。图像是从 ILSVRC 验证集中随机选择的 [55]。red 框表示地面实况框，green 框表示预测框。我们从多层融合的类激活图可以比 Grad-CAM 和 Grad-CAM++ 更精确地定位对象边界框。

表 1. 不同阶段类激活图定位精度的比较。第一行中的 “S” 表示 VGG16 中的 “阶段”。S5-S1 表示 VGG16 中每个阶段的最后一个卷积层。

Method	Metric (%)	S5	S4	S3	S2	S1
CAM	loc1	43.62	18.32	8.87	19.59	13.95
	loc5	53.99	22.70	11.05	23.85	17.27
CAM++	loc1	45.44	41.11	35.33	31.70	31.32
	loc5	56.42	50.97	43.86	39.40	38.90
ScoreCAM	loc1	39.51	33.08	31.15	29.90	29.63
	loc5	49.63	41.63	39.30	37.80	37.46
NormGrad	loc1	38.94	40.85	38.67	32.05	29.94
	loc5	49.19	51.98	49.56	41.37	38.69
LayerCAM	loc1	46.62	44.05	41.83	43.18	43.71
	loc5	57.83	55.02	52.28	53.60	54.34

表 2. 不同阶段融合类激活图定位精度的比较。第一行中的 “S” 表示 VGG16 中的 “阶段”。S5-S1 表示 VGG16 中每个阶段的最后一个卷积层。

Method	Metric (%)	S5	+S4	+S3	+S2	+S1
CAM	loc1	43.62	40.56	40.03	35.87	33.96
	loc5	53.99	50.11	49.47	44.48	42.17
CAM++	loc1	45.44	42.72	37.25	32.51	31.60
	loc5	56.42	52.95	46.19	40.40	39.23
ScoreCAM	loc1	39.51	37.26	31.88	29.94	29.52
	loc5	49.63	46.78	40.19	37.84	37.33
NormGrad	loc1	38.94	36.59	36.45	36.41	36.26
	loc5	49.19	46.02	45.86	45.79	45.59
LayerCAM	loc1	46.62	47.17	47.22	47.24	47.23
	loc5	57.83	58.67	58.72	58.74	58.74

实现细节。为了从类激活图中生成对象边界框，我们直接用最大强度的 15% 的阈值对它们进行二值化，然后找到最大连接段的紧框，如 [92, 56] 中所做的那样。我们选择 VGG-16 中不同阶段的最后一个卷积层来生成类激活图。对于 LayerCAM 从 conv1_2 和 conv2_2 层生成的类激活图，其中具有强值的对象位置倾向于分散在对象周围。因此，延续 [60]，我们应用 GraphCut [5] 来生成连接的分割。此外，当组合来自不同层的类激活图时，来自 VGG16 前三个阶段的类激活图由 Eqn. (9) 缩放。

在表1中，我们首先展示了来自 VGG16 不同阶段的类激活图的定位能力。我们发现 LayerCAM 的定位性能相较 Grad-CAM [56]、Grad-CAM++ [6]、ScoreCAM [70] 和 NormGrad [52] 有大幅度提高，尤其是在浅层。这一事实表明，LayerCAM 可以从浅层获得比 Grad-CAM、Grad-CAM++、ScoreCAM

表 3. γ 的消融实验。

Settings	Metric (%)	1	2	3	4
S5+S4+S3	loc1	47.18	47.22	47.17	47.04
	loc5	58.63	58.72	58.62	58.46
S5+S4+S3+S2	loc1	47.19	47.24	47.17	47.02
	loc5	58.63	58.74	58.63	58.46
S5+S4+S3+S2+S1	loc1	47.20	47.23	47.19	46.97
	loc5	58.65	58.74	58.64	58.39

表 4. 不同尺度函数的消融实验。

Settings	Metric	no scale	$tanh(x)$	$\sqrt[2]{x}$	$tan(x)$
S5+S4+S3	loc1	47.01	47.18	44.52	42.91
	loc5	58.40	58.63	55.62	53.39
S5+S4+S3+S2	loc1	47.00	47.19	44.53	40.07
	loc5	58.39	58.63	55.64	49.96
S5+S4+S3+S2+S1	loc1	46.91	47.20	44.51	38.67
	loc5	58.27	58.65	55.62	48.27

和 NormGrad 更可靠的细粒度目标定位信息。 如图10(b-c) 所示，Grad-CAM 和 Grad-CAM++ 从浅层生成的类激活图无法从背景和其他类别中消除噪声区域。我们在空间维度上为每个位置分配单独权重的 LayerCAM 可以考虑对感兴趣的类别的不同重要性，这样可以在去除背景噪声的同时保持可靠的对象定位信息。

此外，我们还展示了融合不同阶段的类激活图的定位性能。对于来自浅层的类激活图，激活值远低于来自深层的激活值。当我们不使用尺度函数时，融合类激活图的性能不会得到提升，如表4（无尺度）所示。因此，当组合来自不同层的类激活图时，我们首先通过缩放函数从浅层缩放类激活图，其中缩放的图由下式计算

$$\hat{M}^c = \tanh\left(\frac{\gamma * M^c}{\max(M^c)}\right), \tag{9}$$

其中 γ 是缩放因子。然后我们利用一个简单的元素最大值操作来组合来自不同层的图。从表3可以看出，当 γ 设置为 2 时，LayerCAM 取得了最好的定位结果。我们还探索了不同种类的尺度函数，如 表4 所示。当我们使用 $\sqrt[2]{x}$ 缩放函数时，性能变得更糟。这是因为 $\sqrt[2]{x}$ 尺度函数将接近 0 的值放大太多，从而增强了噪声强度。例如，0.01 缩放到 0.1。当我们使用 $tan(x)$ scale 函数时，性能也会变得更糟。$tan(x)$ scale 函数将 1 附近的大值缩放很多，这将抑制归一化后较低值的放大。可以看出，当使用 $tanh(x)$ 尺度函数时，我们可以获得更好的融合结果。

表 5. 不同方法之间定位精度的比较。其他方法的注意力图都是从最后的卷积层生成的。星号 * 表示结果来自这篇文章[56]。

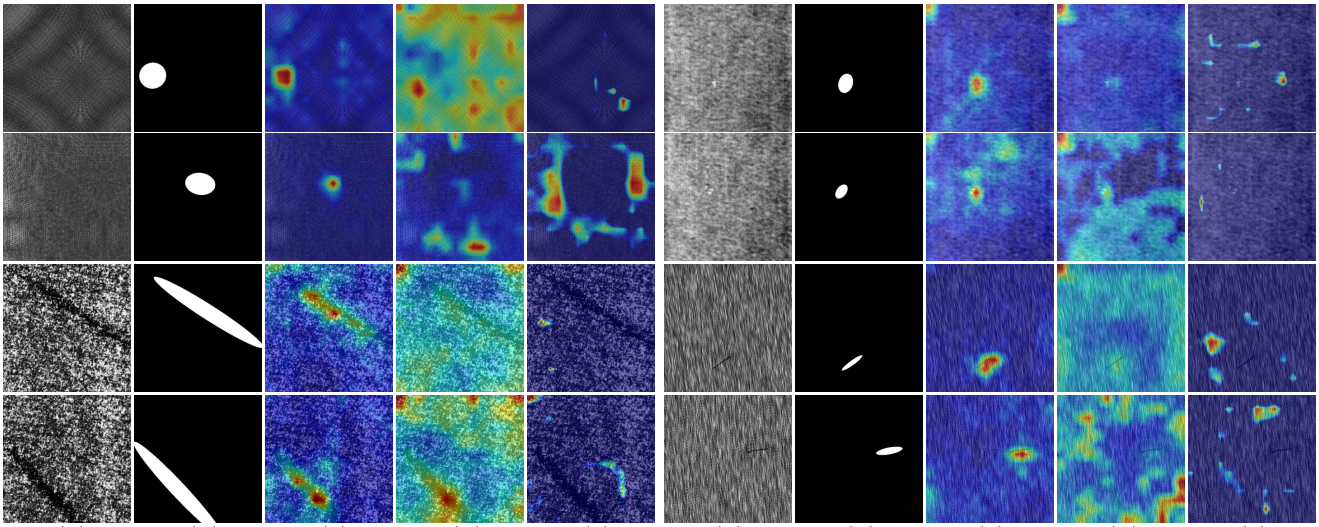
Methods	ACoL	ADL	CAM*	c-MWP*	Ours
loc1 (%)	45.83	44.92	42.80	29.08	47.24
loc5 (%)	59.43	-	54.86	36.96	58.74

如表2所示，来自不同层的类激活图的组合可以逐渐提高定位性能。然而，我们还观察到，当逐渐融合来自浅层的类激活图时，性能增益变得非常小。我们分析了融合地图的定位性能是有限的，因为边界框仅指示一般对象位置。无法测量物体的细粒度细节；例如，如果找到马的耳朵，则边界框不会有太大变化。在第4.4节中，当融合来自浅层的类激活图时，分割结果逐渐增加，例如第 3 阶段的类激活图，这也可以验证融合图的质量。

在表5中，我们展示了不同方法之间定位性能的比较。最右边两列的 attention 方法都是基于原来的 VGG16 模型。最左边的三种定位方法都采用 VGG16 架构，用全局平均池化层代替全连接层。与 c-MWP、CAM、Grad-CAM 和 GradCAM++ 相比，可以看出我们的 LayerCAM 将 $loc1$ 性能分别提高了 18.16%、4.44%、3.62% 和 1.80%。我们的方法也取得了比一些最先进的定位方法 ACoL [89] 和 ADL[9] 更好的结果。它们是专门为解决对象定位任务而设计的。比较表明，我们的 LayerCAM 的类激活图可以提供更可靠的对象定位信息。可视化示例可以在图6 中找到。

4.2. 图像遮挡

对于 LayerCAM 从最终卷积层生成的类激活图，我们进行了 [85]中提出的图像遮挡实验，以验证



(a) (b) (c) (d) (e) (a) (b) (c) (d) (e)

图 7. 工业缺陷的类激活图。(a) 原图 (b) 真值。(c-e) 显示了由 LayerCAM、Gard-CAM++ 和 Grad-CAM 生成的 ResNet50 的 layer3 的类激活图。这些图像是从 DAGM-2007 缺陷数据集 [78] 中随机选择的。我们的类激活图可以比 Grad-CAM 和 Grad-CAM++ 更精确地定位对象边界框。**放大以获得最佳视图**.

表 6. 图像遮挡实验的分类精度比较。Confidence: 表示真实类别的平均预测分数。分数越低，效果越好。

Method	original	Grad-CAM	Grad-CAM++	LayerCAM
Top-1 Acc (%)	68.74	50.36	50.07	48.26
Top-5 Acc (%)	88.57	75.62	75.26	73.43
Confidence (%)	68.64	50.24	49.99	48.12

这些图指示的置信区域的可靠性。图像遮挡实验测试被遮挡的图像区域对最终预测的重要性。如果置信区域很重要，当我们输入被遮挡的图像时，目标类别的预测分数将大大降低。具体来说，在 ILSVRC 验证集上，我们首先选择出被 第4.1节中使用的 VGG16 正确预测的图像。对于这些真实预测的图像，我们用 0.7 的阈值遮挡它们，然后将它们输入到网络中。如 表6 所示，我们分别展示了真实类别的 top-1 分类准确率、top-5 分类准确率和平均预测分数。

表6 展示了图像遮挡实验的性能。LayerCAM 实现了比 Grad-CAM 和 Grad-CAM++ 更低的分类精度，这表明由 LayerCAM 从最终卷积层生成的类激活图可以为目标类别发现更重要的空间对象区域。该实验验证了我们的 LayerCAM 定位的置信区域的可靠性。在对图像进行掩蔽时，我们可以发现 LayerCAM 去除置信区域后比 Grad-CAM 和 Grad-CAM++ 更显着地降低了预测分数。

4.3. 工业表面缺陷定位

对于从 CNN 浅层生成的类激活图，我们利用它们来定位工业图像中各种形状的微小缺陷。我们将此问题视为图像中存在或不存在缺陷的二元分类问题。然后我们使用图像级标签来训练基于 ResNet50 [19] 的分类器。最后，我们应用 LayerCAM 来定位工业图像中的缺陷。

表 7. 不同方法的比较。SegNet 和 RefineNet 是全监督方法，其他方法是弱监督方法。结果中星号 * 表示它们来自原始论文 [12]。

Methods	mIoU (%)	FPS
SegNet [4]	21.95*	17.92*
RefineNet [37]	32.90*	31.05*
Grad-CAM	0.35	60.97
Grad-CAM++	6.46	60.24
LayerCAM	27.26	60.61

实现细节。 我们在 DAGM-2007 缺陷数据集 [78] 上

表 8. 不同层间定位精度的比较。S4-S1 代表 ResNet-50 中的第四层和第一层。

Setting	S4	S3	S2	S1	S4+S3	S3+S2
mIoU (%)	11.59	27.26	19.37	13.10	12.28	24.51

进行了实验，其中包含 3550 个训练图像和 400 个测试图像。

该数据集包含不同纹理表面上的多种类型的缺陷，如 图7(a-b) 所示。我们在这个数据集上训练一个缺陷图像分类器。我们使用 SGD 来优化分类器并将模型训练 15 个 epoch，批大小为 32。初始学习率设置为 0.001，并在第 5 轮和第 10 轮处衰减。在推理时，我们分别将 LayerCAM、Grad-CAM 和 Grad-CAM++ 应用于 ResNet-50 的 layer3 以生成类激活图。生成的图首先被阈值化为二进制掩码。然后我们计算二元掩码和真实掩码之间的 IoU 分数。我们为每种注意力方法寻找最佳阈值并报告它们的最佳性能。我们还测试了不同方法的每秒帧数 (FPS)。运行时间在 NVIDIA RTX 2080Ti 上平均超过 100 次迭代。

我们已经在 表7 中展示了不同方法之间的定量比较。实验结果表明，在抑制背景噪声的同时，我们的方法可以比 Grad-CAM 和 Grad-CAM++ 更准确地定位缺陷。我们还展示了几种使用像素级标签训练的全监督方法 SegNet [4] 和 RefineNet [37]。与它们相比，我们的 LayerCAM 实现了可比的性能，但速度大约是它们的两倍。

此外，我们还在 图7 中展示了不同方法之间的定性比较。与 Grad-CAM 和 Grad-CAM++ 相比，LayerCAM 可以定位各种形状的缺陷，而 Grad-CAM 和 Grad-CAM++ 无法过滤背景上的干扰信息。

在表8中，我们展示了不同层的定位精度。对于工业缺陷定位任务，layer3 的性能优于多层融合的性能。这是因为工业缺陷通常具有小尺寸和各种形状。如图8所示，来自 layer4 的类激活图的低空间分辨率只能粗略地定位缺陷，这不利于特征融合。从 layer2 和 layer1 生成的类激活图定位带有一些噪声

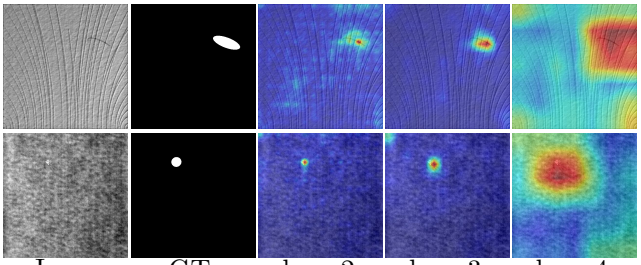


图 8. 来自缺陷定位任务不同层的类激活映射。

的小缺陷区域。因此，对于工业缺陷定位任务，我们仅利用来自 layer3 的类激活图而不是多层融合。

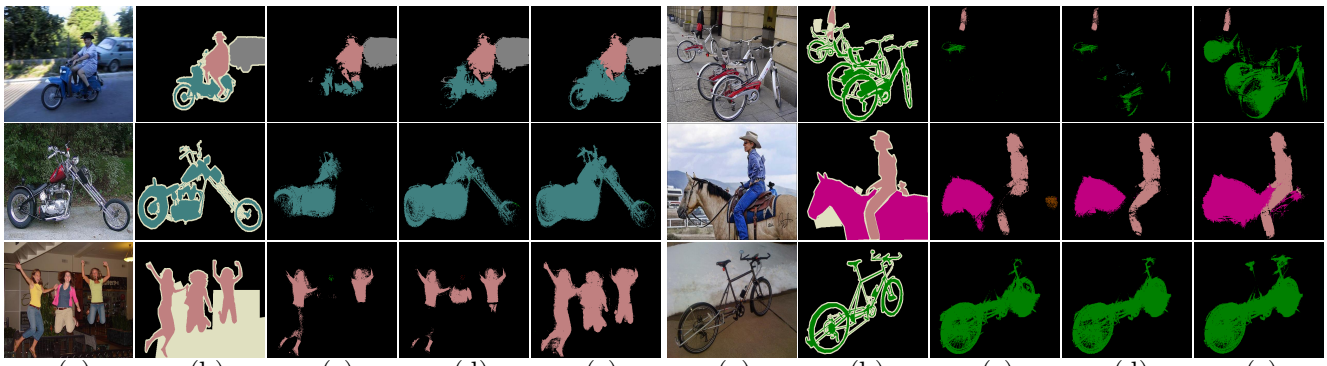
4.4. 弱监督分割

为了进一步测试我们的类激活图的质量，我们将它们应用于需要更多像素精确信息的弱监督语义分割任务。我们利用类激活图和超像素 [3] 来生成伪分割标签。受 [94] 的启发，我们使用类激活图作为查询从超像素中收集对象掩码。我们通过平均每个超像素中的注意力值来计算类别 c 存在的概率，

$$S_c = \left(\frac{1}{|O|} \sum_{j \in O} M_j^c \right), \tag{10}$$

其中 O 表示超像素，M 是类激活图其值归一化到 $[0, 1]$ 的范围内。然后我们在所有目标类别中选择概率最大的，将对应的类别分配给超像素中的所有像素。如果最大概率小于固定的低阈值（在我们的实验中，阈值设置为 0.3），则将超像素中的像素分配给背景类别。 在为每个超像素分配语义类别后，我们利用它们构成伪分割标签来训练分割模型。

实现细节。 我们在流行的 PASCAL VOC 2012 数据集 [13] 上执行分割实验。该数据集包含 20 个语义类和背景。将原始图像拆分为 1464 张训练图像，1449 张验证图像和 1456 个测试图像。按照 [18] 中的设置，我们利用具有 10,582 张图像的增强训练集来训练分割模型，然后在验证和测试集上将我们的方法与 Grad-CAM 和 Grad-CAM++ 进行比较。为了便于比较，我们使用 [92] 中提出的 CNN 分类器。最后一个具有 1000 个通道的全连接 (FC) 层被修改为具有用于 PASCAL VOC 数据集的 20 个通道。我们采用在 ImageNet [55] 上预训练的 VGG16 模



(a) (b) (c) (d) (e) (a) (b) (c) (d) (e)

图 9. 我们的方法产生的分割结果示例：(a) 原图，(b) 真值，(c-e) 使用第 5 阶段，四五阶段组合、三四五阶段组合的类激活图的分割结果。

型来初始化我们的网络并使用交叉熵损失对其进行优化。在推理期间，我们选择从 VGG16 中每个阶段的最后一层生成的类激活图。对于 Grad-CAM 和 Grad-CAM++，我们只利用来自最终卷积层的类激活图，因为来自浅层的图要差得多。

表 9. PASCAL VOC 数据集上的弱监督分割结果。“弱”是指仅具有图像级监督的方法。为了公平比较，我们的方法也基于 Deeplab-LargeFOV 分割模型。

Methods	val (%)	test (%)
Grad-CAM	55.6	56.3
Grad-CAM++	55.5	56.1
LayerCAM (Ours, VGG16)	60.8	61.4
LayerCAM (Ours, ResNet101)	63.0	64.5

我们采用流行的基于 VGG16 [61] 的 Deeplab-LargeFOV [7] 架构作为我们的分割网络。我们还根据 [?] 中的设置，基于 ResNet [19] 训练 Deeplab-LargeFOV [7]。用于训练分割网络的超参数如下：学习率：1e-3；学习率策略：poly，批量大小：10。我们运行 SGD 16000 次。学习率在 12000 次迭代时衰减 10 倍。在推理时，我们使用平均交叉联合（mIoU）度量来评估分割结果。

在 表9 中，我们根据 mIoU 分数报告了我们方法的性能。我们方法的性能优于 Grad-CAM 和 Grad-CAM++ 超过 5 我们还使用类激活图的不同阶段的融合来报告性能，如 表10 所示。我们的方法使用 VGG16 第 5 阶段的类激活图的 mIoU 分数

表 10. 来自不同阶段组合的类激活图时，PASCAL VOC 验证集上的 mIoU 分数的比较。

S5	S4	S3	S2	S1	mIoU (%)
✔					55.6
	✔				55.0
		✔			50.8
			✔		50.5
				✔	46.0
✔	✔				57.1
✔	✔	✔			60.4
✔	✔	✔	✔		60.8
✔	✔	✔	✔	✔	60.2

表 11. DGCN [14] 与不同 CAM 种子的比较。

Setting	val(%)	test (%)
DGCN-CAM	64.0	64.6
DGCN-LayerCAM	67.1	67.6

为 55.6%。我们观察到，当连续将第 4、第 3 和第 2 阶段的类激活图与元素最大操作融合到第 5 阶段时，mIoU 分数逐渐增加（从 55.6% 到 60.8%）。这一事实验证了我们的融合类激活图可以获得更多的对象定位信息，这有利于分割任务。此外，我们还将我们的融合类激活图应用于更高级的弱监督语义方法 DGCN [14]。如 表11 所示，我们可以看到，当用我们的 LayerCAM 种子替换 CAM 种子时，分割结果可以进一步提高约 3 实验结果验证了 LayerCAM 生成的种子比 CAM 生成的种子具有更好的定位能力，这有利于弱监督语义分割方法。 如 图9 所示，

(a) (b) (c) (d) (e)

图 10. 类激活图的可视化。(a) 原图像。(b-d) 分别由 Grad-CAM、Grad-CAM++、LayerCAM 和 LayerCAM-normal 生成的 VGG16 的 “pool2” 层的类激活映射。LayerCAM-normal: 我们使用特征图中每个位置的原始梯度作为其权重。

我们展示了一些定性的分割结果。可以看出，融合 VGG16 不同阶段的类激活图可以逐渐提高分割结果的质量。我们注意到，当将阶段 1 的类激活映射融合到最终映射中时，性能从 60.8% 下降到 60.2%。我们分析，与其他阶段的类激活图相比，第 1 阶段的类激活图缺乏类区分。

4.5. 负梯度的影响

在 Eqn. (6) 中，LayerCAM 利用 ReLU 过滤掉负梯度。在本节中，我们将探讨负梯度的影响。我们首先做实验来研究负梯度对定位能力的影响。如表12所示，负梯度（LayerCAM-normal）的 LayerCAM 实现的定位能力比 LayerCAM 低得多。这一事实证明，LayerCAM 中的负梯度会降低定位能力。此外，我们还测量了 PASCAL VOC 2012 数据集上细粒度位置的 mIoU 分数。我们为 VGG16 的不同阶段生成类激活图，然后通过 0.2 的硬阈值将它们阈值到二进制掩码。我们计算阈值掩码和真实掩码之间的 mIoU 分数。

在 表13 中，我们给出了具有不同设置的 LayerCAM 的 mIoU 分数。可以看出，LayerCAM 使用正梯度作为权重比使用正常梯度（具有负梯度）获得更高的 mIoU 分数，我们还给出了 pool2 层的定性结果 图10(d-e)。可以看出，来自具有负梯度的 LayerCAM 的类激活图丢失了许多对象定位信息。以前的作品 [6, 63, 85] 也表明了正梯度在生成类激活图或显著图方面的重要性。因此，根据经验结果，我们过滤掉负梯度并选择正梯度作为特征图中每个位置的权重。

表 12. 不同阶段类激活图定位精度的比较。第一行中的 “S” 表示 VGG16 中的 “阶段”。S5-S1 表示 VGG16 中每个阶段的最后一个卷积层。

Method	Metric	S5	S4	S3	S2	S1
LayerCAM-normal	loc1	42.09	37.63	34.74	34.12	30.86
	loc5	52.10	46.37	43.09	42.52	39.01
LayerCAM	loc1	46.62	44.05	41.83	43.18	43.71
	loc5	57.83	55.02	52.28	53.60	54.34

表 13. 不同设置下 LayerCAM 的类激活图的比较。LayerCAM-normal: 表示我们使用特征图中每个位置的原始梯度作为其权重。

mIoU(%)	S5	S4	S3	S2	S1
LayerCAM-normal	34.3	22.2	14.4	8.4	4.8
LayerCAM	36.2	35.7	31.5	21.8	11.1

5. Conclusion

在本文中，我们提出了一种注意力方法 LayerCAM，它可以有效地从 CNN 的不同层生成可靠的类激活图。来自深层的类激活图可以定位对象的大致位置，来自浅层的类激活图可以生成细粒度的对象定位信息。来自不同层的类激活图的组合可以找到更多的对象位置，这有利于提高弱监督任务的性能。实验表明，LayerCAM 比当前的注意力方法具有更好的对象定位能力。此外，LayerCAM 易于用于任何现成的基于 CNN 的图像分类器，无需修改网络架构和改变反向传播方式。PyTorch [48] 和 Jittor [23] 版本的源代码都将公开。

致谢

这项研究得到了国家重点研发计划基金号: 2018AAA0100400，国家自然科学基金（61922046），

教育部科技创新项目，中央高校基本科研业务费专项资金（南开大学，编号：63213090）。

参考文献

[1] J. Ahn and S. Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 4981–4990, 2018. 1

[2] N. Araslanov and S. Roth. Single-stage semantic segmentation from image labels. In IEEE Conf. Comput. Vis. Pattern Recog., pages 4253–4262, 2020. 3

[3] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In IEEE Conf. Comput. Vis. Pattern Recog., 2014. 10

[4] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell., 39(12):2481–2495, 2017. 9, 10

[5] Y. Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In Int. Conf. Comput. Vis., volume 1, pages 105–112. IEEE, 2001. 7

[6] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In IEEE Winter Conf. Appl. Comput. Vis., pages 839–847, 2018. 1, 2, 3, 5, 7, 12

[7] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell., 2017. 11

[8] Y. Chen, Y. Lin, M. Yang, and J. Huang. Show, match and segment: Joint weakly supervised learning of semantic matching and object co-segmentation. IEEE Trans. Pattern Anal. Mach. Intell., 2020. 2

[9] J. Choe and H. Shim. Attention-based dropout layer for weakly supervised object localization. In IEEE Conf. Comput. Vis. Pattern Recog., pages 2219–2228, 2019. 2, 8

[10] H. Cholakkal, G. Sun, F. S. Khan, and L. Shao. Object counting and instance segmentation with image-level supervision. In IEEE Conf. Comput. Vis. Pattern Recog., pages 12397–12405, 2019. 2

[11] R. G. Cinbis, J. Verbeek, and C. Schmid. Weakly supervised object localization with multi-fold multiple instance learning. IEEE Trans. Pattern Anal. Mach. Intell., 39(1):189–203, 2016. 3

[12] H. Dong, K. Song, Y. He, J. Xu, Y. Yan, and Q. Meng. Pga-net: Pyramid feature fusion and global context attention network for automated surface defect detection. IEEE TII, 2019. 4, 9

[13] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. Int. J. Comput. Vis., 2015. 6, 10

[14] J. Feng, X. Wang, and W. Liu. Deep graph cut network for weakly-supervised semantic segmentation. Science China Information Sciences, 64(3):130105, 2021. 3, 11

[15] C. Galleguillos, B. Babenko, A. Rabinovich, and S. Belongie. Weakly supervised object localization with stable segmentations. In Eur. Conf. Comput. Vis., pages 193–207, 2008. 3

[16] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr. Res2net: A new multi-scale backbone architecture. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(2):652–662, 2021. 2

[17] R. Gokberk Cinbis, J. Verbeek, and C. Schmid. Multi-fold mil training for weakly supervised object localization. In IEEE Conf. Comput. Vis. Pattern Recog., pages 2409–2416, 2014. 3

[18] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In Int. Conf. Comput. Vis., 2011. 10

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In IEEE Conf. Comput. Vis. Pattern Recog., 2016. 2, 9, 11

[20] S. Hong, D. Yeo, S. Kwak, H. Lee, and B. Han. Weakly supervised semantic segmentation using web-crawled videos. In IEEE Conf. Comput. Vis. Pattern Recog., pages 3626–3635, 2017. 1

[21] Q. Hou, P. Jiang, Y. Wei, and M.-M. Cheng. Self-erasing network for integral object attention. In Adv. Neural Inform. Process. Syst., pages 549–559, 2018. 1, 3

[22] Q. Hou, D. Massiceti, P. K. Dokania, Y. Wei, M.-M. Cheng, and P. H. Torr. Bottom-up top-down cues for weakly-supervised semantic segmentation. In International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition, pages 263–277, 2017. 2

[23] S.-M. Hu, D. Liang, G.-Y. Yang, G.-W. Yang, and W.-Y. Zhou. Jittor: a novel deep learning framework with meta-operators and unified graph execution. Science China Information Sciences, 63(12):1–21, 2020. 12

[24] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In IEEE Conf. Comput. Vis. Pattern Recog., pages 4700–4708, 2017. 2

[25] Z. Huang, X. Wang, J. Wang, W. Liu, and J. Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7014–7023, 2018. 3

[26] P.-T. Jiang, Q. Hou, Y. Cao, M.-M. Cheng, Y. Wei, and H.-K. Xiong. Integral object mining via online attention accumulation. In Int. Conf. Comput. Vis., pages 2070–2079, 2019. 3

[27] Z. Jie, Y. Wei, X. Jin, J. Feng, and W. Liu. Deep self-taught learning for weakly supervised object localization. In IEEE Conf. Comput. Vis. Pattern Recog., pages 1377–1385, 2017. 3

[28] L. Jing, Y. Chen, and Y. Tian. Coarse-to-fine semantic segmentation from image-level labels. IEEE Trans. Image Process., 29:225–236, 2020. 3

[29] V. Kantorov, M. Oquab, M. Cho, and I. Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In Eur. Conf. Comput. Vis., pages 350–365, 2016. 3

[30] D. Kim, D. Cho, D. Yoo, and I. So Kweon. Two-phase learning for weakly supervised object localization. In Int. Conf. Comput. Vis., pages 3534–3543, 2017. 3

[31] D. Li, J. Huang, Y. Li, S. Wang, and M. Yang. Progressive representation adaptation for weakly supervised object localization. IEEE Trans. Pattern Anal. Mach. Intell., 42(6):1424–1438, 2020. 1

[32] D. Li, J.-B. Huang, Y. Li, S. Wang, and M.-H. Yang. Weakly supervised object localization with progressive domain adaptation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 3512–3520, 2016. 3

[33] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu. Tell me where to look: Guided attention inference network. In IEEE Conf. Comput. Vis. Pattern Recog., 2018. 1

[34] X. Li, H. Ma, and X. Luo. Weaklier supervised semantic segmentation with only one image level annotation

per category. IEEE Trans. Image Process., 29:128–141, 2020. 3

[35] X. Li, T. Zhou, J. Li, Y. Zhou, and Z. Zhang. Group-wise semantic mining for weakly supervised semantic segmentation. arXiv preprint arXiv:2012.05007, 2020. 3

[36] G. Lin, A. Milan, C. Shen, and I. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 1925–1934, 2017. 3

[37] G. Lin, A. Milan, C. Shen, and I. Reid. Refinenet: Multi-path refinement networks with identity mappings for high-resolution semantic segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., 2017. 9, 10

[38] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In IEEE Conf. Comput. Vis. Pattern Recog., pages 2117–2125, 2017. 3

[39] J.-J. Liu, Q. Hou, M.-M. Cheng, C. Wang, and J. Feng. Improving convolutional networks with self-calibrated convolutions. In IEEE Conf. Comput. Vis. Pattern Recog., pages 10096–10105, 2020. 2

[40] N. Liu and J. Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In IEEE Conf. Comput. Vis. Pattern Recog., pages 678–686, 2016. 3

[41] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In Eur. Conf. Comput. Vis., pages 21–37, 2016. 3

[42] Y. Liu, P.-T. Jiang, V. Petrosyan, S.-J. Li, J. Bian, L. Zhang, and M.-M. Cheng. Del: Deep embedding learning for efficient image segmentation. In Int. Joint Conf. Artif. Intell., 2018. 3

[43] S. Lu, J. Feng, H. Zhang, J. Liu, and Z. Wu. An estimation method of defect size from mfl image using visual transformation convolutional neural network. IEEE TII, 15(1):213–224, 2019. 4

[44] Z. Lu, Z. Fu, T. Xiang, P. Han, L. Wang, and X. Gao. Learning from weak and noisy labels for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell., 39(3):486–500, 2017. 3

[45] Q. Meng, W. Wang, T. Zhou, J. Shen, Y. Jia, and L. Van Gool. Towards a weakly supervised framework for 3d point cloud object detection and annotation. IEEE Trans. Pattern Anal. Mach. Intell., 2021. 2

[46] Q. Meng, W. Wang, T. Zhou, J. Shen, L. Van Gool, and D. Dai. Weakly supervised 3d object detection from lidar point cloud. In Eur. Conf. Comput. Vis., pages 515–531, 2020. 2

[47] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille. Weakly-and semi-supervised learning of a dcnn for semantic image segmentation. In Int. Conf. Comput. Vis., 2015. 3

[48] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In Adv. Neural Inform. Process. Syst., 2019. 12

[49] D. Pathak, P. Krahenbuhl, and T. Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In ICCV, pages 1796–1804, 2015. 3

[50] B. N. Patro, M. Lunayach, S. Patel, and V. P. Namboodiri. U-cam: Visual explanation using uncertainty based class activation maps. In Int. Conf. Comput. Vis., pages 7444–7453, 2019. 2

[51] J. Pont-Tuset, P. Arbelaez, J. T. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping for image segmentation and object proposal generation. IEEE Trans. Pattern Anal. Mach. Intell., 39(1):128–140, 2016. 3

[52] S.-A. Rebuffi, R. Fong, X. Ji, and A. Vedaldi. There and back again: Revisiting backpropagation saliency methods. In IEEE Conf. Comput. Vis. Pattern Recog., pages 8839–8848, 2020. 3, 7

[53] C. Redondo-Cabrera, M. Baptista-Ríos, and R. J. López-Sastre. Learning to exploit the prior network knowledge for weakly supervised semantic segmentation. IEEE Trans. Image Process., 28(7):3649–3661, 2019. 3

[54] A. Roy and S. Todorovic. Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., 2017. 3

[55] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis., 2015. 6, 7, 10

[56] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explana-

tions from deep networks via gradient-based localization. Int. J. Comput. Vis., 128(2):336–359, 2020. 1, 2, 3, 7, 8

[57] J. Shen, X. Tang, X. Dong, and L. Shao. Visual object tracking by hierarchical attention siamese network. IEEE transactions on cybernetics, 50(7):3068–3080, 2019. 3

[58] Z. Shi, T. M. Hospedales, and T. Xiang. Bayesian joint modelling for object localisation in weakly labelled images. IEEE Trans. Pattern Anal. Mach. Intell., 37(10):1959–1972, 2015. 2

[59] Z. Shi, Y. Yang, T. M. Hospedales, and T. Xiang. Weakly-supervised image annotation and segmentation with objects and attributes. IEEE Trans. Pattern Anal. Mach. Intell., 39(12):2525–2538, 2017. 3

[60] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Int. Conf. Learn. Represent., 2014. 2, 7

[61] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In Int. Conf. Learn. Represent., 2015. 1, 11

[62] K. K. Singh and Y. J. Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In Int. Conf. Comput. Vis., pages 3544–3553. IEEE, 2017. 3

[63] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller. Striving for simplicity: The all convolutional net. In Int. Conf. Learn. Represent. Worksh., 2015. 2, 12

[64] B. Su, H. y. Chen, P. Chen, G. Bian, k. Liu, and W. Liu. Deep learning-based solar-cell manufacturing defect detection with complementary attention network. IEEE TII, 2020. 4

[65] G. Sun, W. Wang, J. Dai, and L. Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In Eur. Conf. Comput. Vis., pages 347–365, 2020. 3

[66] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In IEEE Conf. Comput. Vis. Pattern Recog., pages 2818–2826, 2016. 2

[67] Z. Tang, E. Tian, Y. Wang, L. Wang, and T. Yang. Non-destructive defect detection in castings by using spatial attention bilinear convolutional neural network. IEEE TII, 2020. 4

[68] E. W. Teh, M. Rochan, and Y. Wang. Attention networks for weakly supervised object localization. In Brit. Mach. Vis. Conf., pages 1–11, 2016. 3

[69] F. Wan, C. Liu, W. Ke, X. Ji, J. Jiao, and Q. Ye. C-mil: Continuation multiple instance learning for weakly supervised object detection. In IEEE Conf. Comput. Vis. Pattern Recog., pages 2199–2208, 2019. 3

[70] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In IEEE Conf. Comput. Vis. Pattern Recog. Worksh., pages 24–25, 2020. 2, 7

[71] W. Wang, J. Shen, and H. Ling. A deep network solution for attention and aesthetics aware photo cropping. IEEE Trans. Pattern Anal. Mach. Intell., 41(7):1531–1544, 2018. 2

[72] W. Wang, J. Shen, X. Lu, S. C. Hoi, and H. Ling. Paying attention to video object pattern understanding. IEEE Trans. Pattern Anal. Mach. Intell., 2020. 2

[73] W. Wang, S. Zhao, J. Shen, S. C. Hoi, and A. Borji. Salient object detection with pyramid attention and salient edges. In IEEE Conf. Comput. Vis. Pattern Recog., pages 1448–1457, 2019. 3

[74] W. Wang, T. Zhou, S. Qi, J. Shen, and S.-C. Zhu. Hierarchical human semantic parsing with comprehensive part-relation modeling. IEEE Trans. Pattern Anal. Mach. Intell., 2021. 3

[75] W. Wang, H. Zhu, J. Dai, Y. Pang, J. Shen, and L. Shao. Hierarchical human parsing with typed part-relation reasoning. In IEEE Conf. Comput. Vis. Pattern Recog., pages 8929–8939, 2020. 3

[76] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In IEEE Conf. Comput. Vis. Pattern Recog., 2017. 3

[77] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., 2018. 3

[78] M. Wieler and T. Hahn. Weakly supervised learning for industrial optical inspection, 2007. 9

[79] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In IEEE Conf. Comput. Vis. Pattern Recog., pages 1492–1500, 2017. 2

[80] S. Xie and Z. Tu. Holistically-nested edge detection. In Int. Conf. Comput. Vis., pages 1395–1403, 2015. 3

[81] W. Xu, Y. Wu, W. Ma, and G. Wang. Adaptively denoising proposal collection for weakly supervised object localization. Neural Processing Letters, 51(1):993–1006, 2020. 3

[82] H. Xue, C. Liu, F. Wan, J. Jiao, X. Ji, and Q. Ye. Danet: Divergent activation for weakly supervised object localization. In Int. Conf. Comput. Vis., pages 6589–6598, 2019. 3

[83] S. Yang, Y. Kim, Y. Kim, and C. Kim. Combinational class activation maps for weakly supervised object localization. In IEEE Winter Conf. Appl. Comput. Vis., pages 2941–2949, 2020. 2

[84] S. Zagoruyko and N. Komodakis. Wide residual networks. In Brit. Mach. Vis. Conf., 2016. 2

[85] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In Eur. Conf. Comput. Vis., pages 818–833, 2014. 2, 8, 12

[86] J. Zhang, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff. Top-down neural attention by excitation backprop. In Eur. Conf. Comput. Vis., 2016. 2

[87] L. Zhang, Y. Gao, Y. Xia, K. Lu, J. Shen, and R. Ji. Representative discovery of structure cues for weakly-supervised image segmentation. IEEE Trans. Multimedia, 16(2):470–479, 2014. 3

[88] L. Zhang, Y. Yang, Y. Gao, Y. Yu, C. Wang, and X. Li. A probabilistic associative model for segmenting weakly supervised images. IEEE Trans. Image Process., 23(9):4150–4159, 2014. 3

[89] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. Huang. Adversarial complementary learning for weakly supervised object localization. In IEEE Conf. Comput. Vis. Pattern Recog., 2018. 1, 3, 8

[90] X. Zhang, Y. Wei, G. Kang, Y. Yang, and T. Huang. Self-produced guidance for weakly-supervised object localization. In Eur. Conf. Comput. Vis., pages 597–613, 2018. 3

[91] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In IEEE Conf. Comput. Vis. Pattern Recog., 2017. 3

[92] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In IEEE Conf. Comput. Vis. Pattern Recog., 2016. 1, 2, 3, 7, 10

[93] B. Zhou, Y. Sun, D. Bau, and A. Torralba. Interpretable basis decomposition for visual explanation. In Eur. Conf. Comput. Vis., pages 119–134, 2018. 2

[94] Y. Zhou, Y. Zhu, Q. Ye, Q. Qiu, and J. Jiao. Weakly supervised instance segmentation using class peak response. In IEEE Conf. Comput. Vis. Pattern Recog., pages 3791–3800, 2018. 3, 10

[95] Y. Zhu, Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao. Soft proposal networks for weakly supervised object localization. In Int. Conf. Comput. Vis., pages 1841–1850, 2017. 3