

铁路不是火车：显著性作为伪像素监督 用于弱监督语义分割

Seungho Lee*

延世大学

seungholee@yonsei.ac.kr

Minhyun Lee*

延世大学

lmh315@yonsei.ac.kr

Jongwuk Lee

成均馆大学

jongwuklee@skku.edu

Hyunjung Shim†

延世大学

kateshim@yonsei.ac.kr

Abstract

现有使用图像级弱监督的弱监督语义分割 (WSSS) 研究存在几个局限性：稀疏的对象覆盖、不准确的对象边界以及来自非目标对象的共现像素。为了克服这些挑战，我们提出了一种新颖的框架，即显式伪像素监督 (EPS)，通过结合两种弱监督从像素级反馈中学习；图像级标签通过定位图提供对象身份，而来自现成显著性检测模型的显著性图提供丰富的边界。我们设计了一种联合训练策略，以充分利用两种信息之间的互补关系。我们的方法可以获得准确的对象边界并丢弃共现像素，从而显著提高伪掩码的质量。实验结果表明，所提出的方法通过解决 WSSS 的关键挑战显著优于现有方法，并在 PASCAL VOC 2012 和 MS COCO 2014 数据集上实现了新的最先进性能。代码可在 <https://github.com/halbielee/EPS> 获取。

1. 介绍

弱监督语义分割 (WSSS) 利用弱监督（例如，图像级标签 [36, 37]、涂鸦 [29] 或边界框 [22]）并旨在实现与需要像素级标签的全监督模型相竞争的性能。大多数现有研究采用图像级标签作为分割模型的弱监督。WSSS 的整体流程包括两个阶段。首先，使用图像分类器为目标对象生成伪掩码。然后，使用伪掩码作为监督训练分割模型。生成伪掩码的流行技术是类激活映射 (CAM) [52]，它提供与其图

*表示同等贡献。

†Hyunjung Shim 是通讯作者。

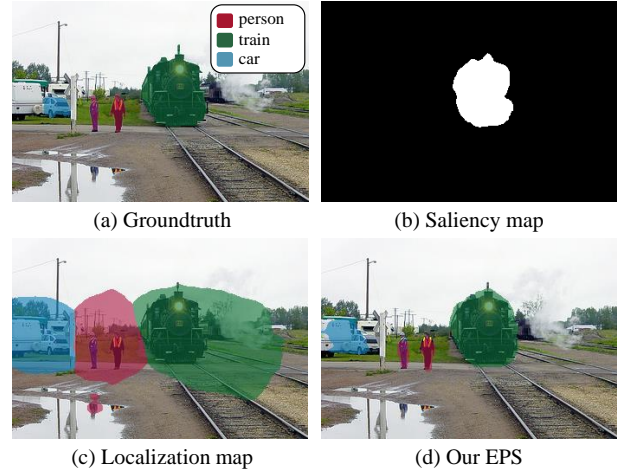


图 1. 利用显著性图和定位图进行 WSSS 的激励示例。(a) 真实值，(b) 通过 PFAN [51]生成的显著性图，(c) 通过 CAM [52]生成的定位图，以及 (d) 我们的 EPS 利用显著性图和定位图来训练分类器。请注意，显著性图无法捕捉人和车，而我们的结果可以正确恢复它们，并且定位图过度捕捉了两个对象。

像级标签对应的对象定位图。由于全监督（即像素级注释）和弱监督（即图像级标签）语义分割之间的监督差距，WSSS 面临以下关键挑战：1) 定位图仅捕获目标对象的一小部分 [52]，2) 它遭受对象边界不匹配的困扰 [23]，以及 3) 它几乎无法将共现像素与目标对象分开（例如，火车上的铁路）[25]。

为了解决这些问题，现有的研究可以分为三个支柱。第一种方法通过擦除像素 [9, 23, 28]、集成得分图 [21, 27]或使用自监督信号 [41]来扩展对象覆盖范围以捕捉对象的全部范围。然而，由于没有线索来指导对象的形状，它们无法确定目标对象的准确边界。第二种方法专注于改进伪掩码的对象边

界 [13, 32]。由于它们有效地学习了对象边界，因此它们自然地将伪掩码扩展到边界。然而，它们仍然无法区分非目标对象与目标对象的重合像素。这是因为前景和背景之间的强相关性（即，共现）几乎无法与归纳偏差（即，观察目标对象及其重合像素的频率）区分开来，如 [10]所示。最后，第三种方法旨在通过使用额外的真实掩码 [24]或显著性图 [35, 47]来缓解共现问题。然而，[24, 28]需要强像素级注释，这与弱监督学习范式相去甚远。[35]对显著性图的错误很敏感。此外，[47]未能覆盖对象的全部范围，并且存在边界不匹配的问题。

在本文中，我们的目标是通过充分利用定位图（即，使用图像级标签训练的图像分类器的 CAM）和显著性图（即，现成的显著性检测模型的输出 [18, 34, 51]）来克服 WSSS 的三个挑战。我们专注于定位图和显著性图之间的互补关系。如图 1所示，定位图可以区分不同的对象，但不能有效地分离它们的边界。相反，虽然显著性图提供了丰富的边界信息，但它并未揭示对象的身份。从这个意义上说，我们认为使用这两种互补信息的方法可以解决 WSSS 的性能瓶颈。

为此，我们提出了一种新的 WSSS 框架，称为显式伪像素监督（EPS）。为了充分利用显著性图（即，前景和背景），我们设计了一个分类器来预测 $C + 1$ 类，包括 C 个目标类和背景类。我们利用 C 个定位图和背景定位图来估计显著性图。然后，显著性损失被定义为显著性图和我们估计的显著性图之间的像素级差异。通过引入显著性损失，模型可以通过所有类的伪像素反馈进行监督。我们还使用多标签分类损失来预测图像级标签。因此，我们训练分类器以优化显著性损失和多标签分类损失，协同优化背景和前景像素的预测——我们发现我们的策略可以改进显著性图（第 3.3节和图 3）和伪掩码（第 5.1节和图 4）。

我们强调，由于显著性损失通过伪像素反馈惩罚边界不匹配，它可以强制我们的方法学习对象的准确边界。作为副产品，我们还可以通过将地图扩展到边界来捕捉整个对象。因为显著性损失有助于将前景（例如，火车）与背景分开，我们的方法可以

将共现像素（例如，铁路）分配给背景类。实验结果表明，我们的 EPS 在 PASCAL VOC 2012 和 MS COCO 2014 数据集上取得了显著的分割性能，记录了新的最先进的准确性。

2. 相关工作

弱监督语义分割。 WSSS 的一般流程是从分类网络生成伪掩码，并使用伪掩码作为监督来训练分割网络。由于图像级标签中边界信息的稀缺，许多现有方法存在伪掩码不准确的问题。为了解决这个问题，交叉图像亲和性 [15]、知识图谱 [31] 和对比优化 [38, 50] 被用来提高伪掩码的质量。[5] 提出了一种自监督任务来发现子类别，以强制分类器改进 CAM。[1, 2] 通过计算像素之间的亲和性隐式利用边界信息。[49] 专注于生成可靠的像素级注释，并设计了一个端到端网络来生成分割图。[20, 25] 通过利用边界损失来训练分割网络。最近，[3] 使用了一个基于单分割的模型，并采用自监督训练方案。[14] 通过利用多个不完整的伪掩码来关注分割网络的鲁棒性。

基于显著性引导的语义分割。 显著性检测（SD）方法通过具有像素级注释的外部显著性数据集 [18, 46, 51] 或图像级注释 [39] 生成区分图像中前景和背景的显著性图。许多 WSSS 方法 [15, 20, 27, 28, 42, 44] 利用显著性图作为伪掩码的背景线索。[43] 利用显著性图作为单对象图像的完整监督。[16] 使用实例级显著性图来学习对象的相似性图。[6, 40, 47] 将显著性图与特定类别的注意力线索结合起来生成可靠的伪掩码。[48] 使用单一网络联合解决 WSSS 和 SD，以提高两项任务的性能。我们的 EPS 可以归类为显著性引导的方法，但在以下原因中明显区别于其他方法。大多数现有方法将显著性图作为伪掩码的一部分或作为细化分类器中间特征的隐式指导。相反，我们的方法利用显著性图作为定位图的伪像素反馈。尽管 [48] 在利用两种互补信息的意义上与我们的方法最为相似，但他们既没有解决共现问题，也没有处理噪声显著性图问题。

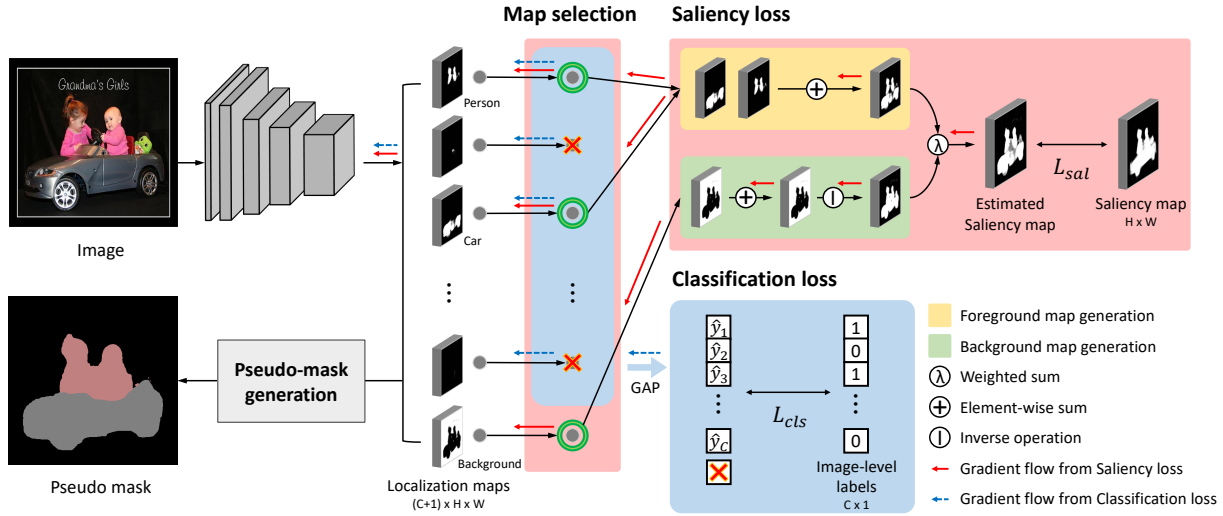


图 2. 我们 EPS 的整体框架。 $C+1$ 个定位图是从一个主干网络生成的。实际的显著性图是从现成的显著性检测模型生成的。某些目标标签的定位图被选择性地用于生成估计的显著性图（第 3.2 节）。整体框架与显著性损失和分类损失共同训练（第 3.3 节）。

3. 提出的方法

在本节中，我们提出了一种新的弱监督语义分割（WSSS）框架，称为显式伪像素监督（EPS）。考虑到 WSSS 的两个阶段，第一阶段是生成伪掩码，第二阶段是训练分割模型。在这里，我们的主要贡献是生成准确的伪掩码。遵循 WSSS 的惯例 [13, 21, 27, 28, 41, 42]，我们然后训练一个分割模型，其中第一阶段生成的伪掩码用作监督。

3.1. 动机

我们 EPS 的关键见解是充分利用两种互补信息，即来自定位图的对象身份和来自显著性图的边界信息。为此，我们利用显著性图作为定位图的伪像素反馈，针对目标标签和背景。我们设计了一个具有额外背景类别的分类器，导致预测总共 $C+1$ 个类别，如图 2 所示。使用该分类器，我们可以学习 $C+1$ 个定位图，即 C 个目标标签的定位图和一个背景定位图。然后，我们解释了 EPS 如何解决 WSSS 中的边界不匹配和共现问题。为了处理边界不匹配问题，我们从 C 定位图中估计前景图，并将其与显著性图的前景进行匹配。通过这种方式，目标标签的定位图可以从显著性图中接收到伪像素反馈，从而改善对象的边界。为了减轻非目标对象的共现像素，我们还将背景的定位图与显著性图进行匹配。

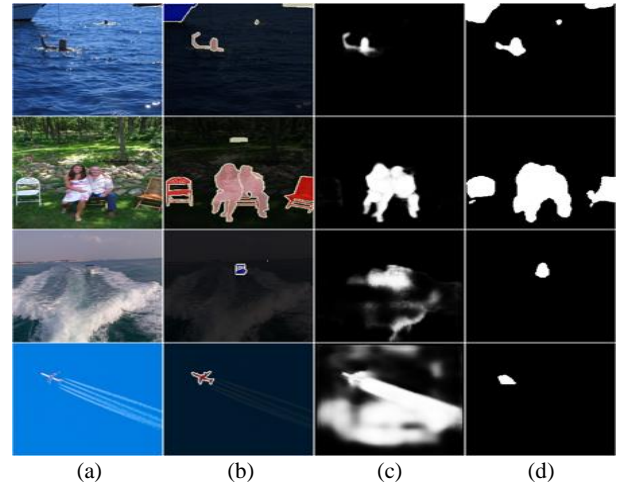


图 3. PASCAL VOC 2012 上估计显著性图的定性示例。(a) 输入图像，(b) 真实值，(c) 来自 [51] 的显著性图和 (d) 我们估计的显著性图。

由于背景的定位图也从显著性图中接收到伪像素反馈，共现像素可以成功地分配给背景；非目标对象的共现像素大多与背景重叠。这就是为什么我们的方法可以将共现像素与目标对象分离的原因。最后，EPS 的目标函数由两部分组成：通过显著性图的显著性损失 \mathcal{L}_{sal} （在图 2 中用红色框/箭头标记）和通过图像级标签的多标签分类损失 \mathcal{L}_{cls} （在图 2 中用蓝色框/箭头标记）。通过联合训练这两个目标，我们可以将定位图和显著性图与互补信息协同起来——我们观察到，通过我们的联合训练策略，彼此的噪

声和缺失信息得到了补充，如图 3 所示。例如，从现成模型 [18, 34, 51] 获得的原始显著性图存在缺失和噪声信息。另一方面，我们的结果成功恢复了缺失的对象（例如，船或椅子）并去除了噪声（例如，水泡或尾迹），这显然比原始显著性图更好。因此，EPS 可以捕捉到更准确的对象边界，并将共现像素与目标对象分离。这些优势带来了显著的性能提升；表 6 报告称，EPS 在分割准确性方面显著优于现有模型，提升幅度高达 3.8–10.6

3.2. 显式伪像素监督

我们解释了如何利用显著性图进行伪像素监督。显著性图的关键优势在于提供对象轮廓，可以更好地揭示对象边界。为了利用这一特性，我们将显著性图与两种情况进行匹配：前景和背景。为了使类级定位图与显著性图具有可比性，我们合并目标标签的定位图并生成前景图， $M_{fg} \in \mathbb{R}^{H \times W}$ 。我们还可以通过将背景图进行反转来表示前景图，背景图是背景标签的定位图 $M_{bg} \in \mathbb{R}^{H \times W}$ 。（稍后，我们将解释如何细化前景图以解决噪声显著性图的问题。）

具体来说，我们使用 M_{fg} 和 M_{bg} 估计显著性图 \hat{M}_s ，如下所示：

$$\hat{M}_s = \lambda M_{fg} + (1 - \lambda)(1 - M_{bg}), \quad (1)$$

其中 $\lambda \in [0, 1]$ 是一个超参数，用于调整前景图和背景图反转的加权和。（在我们的实验中，默认将 λ 设置为 0.5，关于 λ 的附加消融研究见补充材料。）然后，我们将显著性损失 \mathcal{L}_{sal} 定义为我们估计的显著性图与实际显著性图之间的像素级差异之和。（ \mathcal{L}_{sal} 的正式定义在第 3.3 节中给出。）

值得注意的是，使用预训练模型被视为弱监督学习，因此利用显著性图已被广泛接受为 WSSS 中的常见做法。尽管其受欢迎程度很高，但采用完全监督的显著性检测模型可能存在争议，因为它们使用来自不同数据集的像素级注释。在本文中，我们研究了不同显著性检测方法的效果；1) 无监督和 2) 完全监督的显著性检测模型（见第 5.3 节），并通过实验证明我们的方法使用其中任何一种都优于所有其他方法 [13, 21, 40, 43, 47] 使用完全监督的显著性模型。虽然现有方法在充分利用显著性图方面有限，

但我们的方法将显著性图作为伪像素监督，并将其作为边界和共现像素的线索加以利用。**处理显著性偏差的地图选择。**之前，我们假设前景图可以是目标标签的定位图的并集；背景图可以是背景标签的定位图。然而，这种简单的选择规则可能与现成模型计算的显著图不兼容。例如，来自 [51] 的显著图通常会忽略一些物体作为显著物体（例如，图 1 中火车旁的小人物）。这种系统性错误是不可避免的，因为显著性模型学习了不同数据集的统计信息。除非考虑到这种错误，否则相同的错误可能会传播到我们的模型中并导致性能下降。

为了应对系统性错误，我们开发了一种有效的策略，使用定位图和显著图之间的重叠率。具体来说，如果第 i 个定位图 M_i 与显著图的重叠率超过 $\tau\%$ ，则将其分配给前景，否则分配给背景。形式上，前景和背景图的计算如下：

$$\begin{aligned} M_{fg} &= \sum_{i=1}^C y_i \cdot M_i \cdot \mathbb{1}[\mathcal{O}(M_i, M_s) > \tau], \\ M_{bg} &= \sum_{i=1}^C y_i \cdot M_i \cdot \mathbb{1}[\mathcal{O}(M_i, M_s) \leq \tau] + M_{C+1}, \end{aligned} \quad (2)$$

其中 $y \in \mathbb{R}^C$ 是二进制图像级标签， $\mathcal{O}(M_i, M_s)$ 是计算 M_i 和 M_s 之间重叠率的函数。为此，我们首先将定位图和显著图二值化：对于像素 p ， $B_k(p) = 1$ 如果 $M_k(p) > 0.5$ ；否则 $B_k(p) = 0$ 。 B_i 和 B_s 分别是对应于 M_i 和 M_s 的二值化图。然后我们计算 M_i 和 M_s 之间的重叠率，即 $\mathcal{O}(M_i, M_s) = |B_i \cap B_s| / |B_i|$ 。我们将 $\tau = 0.4$ 设置为不考虑数据集和骨干模型。在补充材料中，我们展示了我们的方法对 τ 的选择具有鲁棒性（即， τ 在 $[0.3, 0.5]$ 范围内表现出相当的性能）。

我们将背景标签的单一一定位图与未选择为前景的定位图结合起来，而不是单一的背景标签定位图。尽管简单，但我们可以绕过显著图的错误，并有效地训练显著图中被忽略的一些物体。（在表 3 中，我们报告了所提出策略克服显著图错误的有效性。）

3.3. 联合训练过程

使用显著图和图像级标签，EPS 的整体训练目标由两个部分组成，显著性损失 \mathcal{L}_{sal} 和分类损失

\mathcal{L}_{cls} 。首先，显著性损失 \mathcal{L}_{sal} 通过测量实际显著图 M_s 和估计显著图 \hat{M}_s 之间的平均像素级距离来制定。

$$\mathcal{L}_{sal} = \frac{1}{H \cdot W} \|M_s - \hat{M}_s\|^2, \quad (3)$$

其中 M_s 是从现成的显著性检测模型——PFAN [51] 训练于 DUTS 数据集 [39] 中获得的。请注意，我们的方法始终优于所有先前的艺术，无论显著性检测模型如何。

接下来，分类损失通过图像级标签 y 和其预测 $\hat{y} \in \mathbb{R}^C$ 之间的多标签软边缘损失来计算，后者是对每个目标类的定位图进行全局平均池化的结果。

$$\mathcal{L}_{cls} = -\frac{1}{C} \sum_{i=1}^C y_i \log \sigma(\hat{y}_i) + (1 - y_i) \log (1 - \sigma(\hat{y}_i)), \quad (4)$$

其中 $\sigma(\cdot)$ 是 sigmoid 函数。最后，总训练损失是多标签分类损失和显著性损失的总和，即 $\mathcal{L}_{total} = \mathcal{L}_{cls} + \mathcal{L}_{sal}$ 。如图 2 所示， \mathcal{L}_{sal} 涉及更新 $C + 1$ 类的参数，包括目标对象和背景。同时， \mathcal{L}_{cls} 仅评估 C 类的标签预测，不包括背景类——来自 \mathcal{L}_{cls} 的梯度不会流入背景类。然而，背景类的预测可以被 \mathcal{L}_{cls} 隐式影响，因为它监督分类器的训练。

4. 实验设置

数据集。我们在两个流行的基准数据集上进行实证研究，PASCAL VOC 2012 [12] 和 MS COCO 2014 [30]。PASCAL VOC 2012 包含 21 个类（即 20 个对象和背景），分别有 1,464、1,449 和 1,456 张用于训练、验证和测试集的图像。按照语义分割的常见做法，我们使用包含 10,582 张图像的增强训练集 [17]。接下来，COCO 2014 包含 81 个类，包括一个背景，分别有 82,081 和 40,137 张用于训练和验证的图像，其中不包含目标类的图像被排除，如 [9] 所做。由于某些对象的真实分割标签相互重叠，我们采用 COCO-Stuff [4] 的真实分割标签，解决了同一 COCO 数据集上的重叠问题。

评估协议。我们在 PASCAL VOC 2012 的验证集和测试集以及 COCO 2014 的验证集上验证我们的方法。PASCAL VOC 2012 测试集的评估结果来自官方 PASCAL VOC 评估服务器。此外，我们采用平均交并比（mIoU）来衡量分割模型的准确性。

方法	召回率 (%)	精确率 (%)	F1-分数 (%)
CAM [52] _{CVPR'16}	22.3	35.8	27.5
SEAM [41] _{CVPR'20}	40.2	45.0	42.5
BES [32] _{ECCV'20}	45.5	46.4	45.9
我们的 EPS	60.0	73.1	65.9

表 1. 在 SBD trainval 集上评估的边界准确性。请注意，BES 的结果是从 [32] 中提出的边界预测网络测量的。

实现细节。我们选择 ResNet38 [45] 作为我们方法的主干网络，输出步幅为 8。所有主干模型均在 ImageNet [11] 上进行预训练。我们使用批量大小为 8 的 SGD 优化器。我们的方法训练到 20k 次迭代，学习率为 0.01（最后一个卷积层为 0.1）。对于数据增强，我们使用随机缩放、随机翻转和随机裁剪到 448×448 。对于分割网络，我们采用 DeepLab-LargeFOV (V1) [7] 和 DeepLab-ASPP (V2) [8]，以及 VGG16 和 ResNet101 作为其主干网络。具体来说，我们使用四个分割网络：基于 VGG16 的 DeepLab-V1 和 DeepLab-V2，基于 ResNet101 的 DeepLab-V1 和 DeepLab-V2。更详细的设置在补充材料中。

5. 实验结果

5.1. 处理边界和共现问题

边界不匹配问题。为了验证伪掩码的边界，我们与最先进的方法 [32, 41, 52] 比较边界的质量。我们利用 SBD [17]，它提供了边界注释和 PASCAL VOC 2011 的边界基准。如 [32] 所做，边界的质量通过计算伪掩码的拉普拉斯边缘检测器的边缘以类无关的方式进行评估。然后，通过测量召回率、精确率和 F1 分数，比较预测和真实边界来评估边界质量。表 1 报告了我们的方法在所有三个指标上大大优于其他方法。图 4 中的定性示例显示，我们的方法可以比所有其他方法捕捉到更准确的边界。

共现问题。如几项研究中讨论的 [20, 25, 28, 35]，我们观察到在 PASCAL VOC 2012 中，一些背景类经常与目标对象一起出现。我们通过使用 PASCAL-CONTEXT 数据集 [33] 进行定量分析，该数据集为整个场景提供像素级注释（例如，水和 铁路）。我们

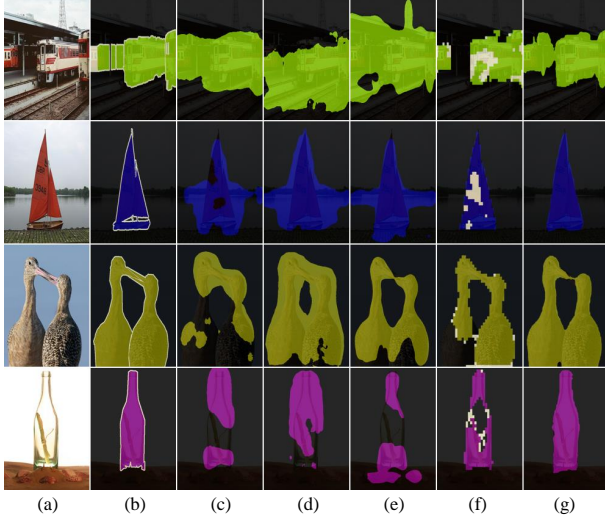


图 4. PASCAL VOC 2012 上伪掩码的定性比较。(a) 输入图像, (b) 真实值, (c) CAM, (d) SEAM, (e) ICD, (f) SGAN 和 (g) 我们的 EPS。

选择了三个共现对; 船和 水, 火车和 铁路, 以及 火车和 平台。我们比较目标类的 IoU 和目标类与其共现类之间的混淆率。混淆率衡量共现类被错误预测为目标类的程度。混淆率 $m_{k,c}$ 通过 $m_{k,c} = FP_{k,c}/TP_c$ 计算, 其中 $FP_{k,c}$ 是共现类 k 被错误分类为目标类 c 的像素数, TP_c 是目标类 c 的真阳性像素数。关于共现问题的更详细分析在补充材料中。表 2 报告了 EPS 的混淆率始终低于其他方法。SGAN [47] 的混淆率与我们的方法非常相似, 但我们的方法在 IoU 方面更准确地捕捉目标类别。有趣的是, SEAM 显示出较高的混淆率, 甚至比 CAM 更差。这是因为 SEAM [41] 通过应用自监督训练来学习覆盖目标对象的全部范围, 这很容易被目标对象的重合像素所迷惑。同时, CAM 只捕捉目标对象中最具辨别力的区域, 而不覆盖较不具辨别力的部分, 例如重合类。我们也可以在图 4 中观察到这种现象。

5.2. 地图选择策略的效果

我们评估了我们的地图选择策略在减轻显著性图误差方面的有效性。我们将三种不同的地图选择策略与不使用地图选择模块的基线进行比较。作为简单策略, 前景图是所有对象定位图的并集; 背景图等于背景类的定位图 (即简单策略)。接下来, 我们遵循简单策略, 但有以下例外。几个预定类 (例

方法	船 w/ 水	火车 w/ 铁路	火车 w/ 站台
CAM [52] _{CVPR'16}	0.74 (33.1)	0.11 (52.9)	0.09 (49.6)
SEAM [41] _{CVPR'20}	1.13 (30.7)	0.24 (48.6)	0.20 (45.5)
ICD [13] _{CVPR'20}	0.47 (41.4)	0.11 (56.7)	0.09 (49.2)
SGAN [47] _{ACCESS'20}	0.10 (42.3)	0.02 (48.8)	0.01 (36.3)
我们的 EPS	0.10 (55.0)	0.02 (78.1)	0.01 (73.0)

表 2. 与处理共现问题的代表性现有方法的比较。每个条目是 $m_{k,c}$ 在 蓝色 中 (越低越好), 括号中的 IoU (越高越好)。

	基线	简单	预定义	我们的自适应
mIoU	66.1	66.5	67.9	69.4

表 3. 地图选择策略的效果。使用不同地图选择策略的伪掩码的准确性在 PASCAL VOC 2012 训练集上进行评估。

方法	无 精炼	有 CRF [26]	有 AffinityNet [2]
CAM [52] _{CVPR'16}	48.0	-	58.1
SEAM [41] _{CVPR'20}	55.4	56.8	63.6
ICD [32] _{CVPR'20*}	59.9	62.2	-
SGAN [47] _{ACCESS'20*}	62.8	-	-
我们的 EPS	69.4	71.4	71.6

表 4. 在 PASCAL VOC 2012 训练集上评估的伪掩码的准确性 (mIoU)。注意, * 表示忽略低置信度像素; 其他方法使用所有像素进行评估。

如, 沙发、椅子和餐桌) 的定位图被分配到背景图 (即预定义类策略)。最后, 所提出的选择方法利用定位图和显著性图之间的重叠率, 如第 3.2 节所述 (即我们的自适应策略)。

表 3 显示了我们的自适应策略可以有效处理显著性图的系统偏差。简单策略意味着在从定位图生成估计显著性图时没有考虑偏差。在这种情况下, 伪掩码的性能下降, 尤其是在沙发、椅子或餐桌类上。使用预定义类的性能表明, 通过忽略显著性图中缺失的类可以减轻偏差。然而, 由于需要人工观察者进行手动选择, 因此不太实用, 无法为每张图像做出最佳决策。同时, 我们的自适应策略可以自动处理偏差, 并为给定的显著性图做出更有效的决策。



图 5. PASCAL VOC 2012 上分割结果的定性示例。(a) 输入图像, (b) 真实值和 (c) 我们的 EPS。

5.3. 与最先进方法的比较

伪掩码的准确性。我们通过聚合不同尺度图像的预测结果采用多尺度推理,这是一种常用的实践,利用用于 [2, 41]。然后,我们通过将我们的 EPS 与基线 CAM [52] 和三种最先进的方法进行比较来评估训练集中的伪掩码的准确性,即 SEAM [41]、ICD [13] 和 SGAN [47]。在这里,测量训练集中伪掩码的准确性是 WSSS 中的常见协议,因为训练集的伪掩码用于监督分割模型。表 4 总结了伪掩码的准确性,并表明我们的方法明显优于所有现有方法,差距为 7–21%。图 4 可视化了伪掩码的定性示例,确认我们的方法显著改善了对象边界,并在伪掩码质量方面显著优于三种最先进的方法。我们的方法可以捕捉对象的精确边界 (第二行),因此自然覆盖对象的全部范围 (第三行),并减轻重合像素 (第一行)。更多示例和失败案例在补充材料中提供。

分割图的准确性。以前的方法 [2, 13, 41] 生成伪掩码并使用 CRF 后处理算法 [26] 或亲和网络 [2] 对其进行细化。同时,如表 4 所示,我们生成的伪掩码足够准确,因此我们在没有任何额外细化的情况下训练分割网络。我们在 Pascal VOC 2012 数据集的四个分割网络上广泛评估并精确比较我们的方法与其他方法。

我们的方法在分割网络方面表现显著优于其他方法。表 5 报告了在相同的 VGG16 主干网络下,我们的方法比其他方法更准确。此外,我们在 VGG16 上的结果与基于更强大主干网络 (i.e. 表 6

方法	分割	支持	验证	测试
SEC [25]ECCV'16	V1	I.	50.7	51.7
AffinityNet [2]CVPR'18	V1	I.	58.4	60.5
ICD [13]CVPR'20	V1	I.	61.2	60.9
BES [32]ECCV'20	V1	I.	60.1	61.1
GAIN [28]CVPR'18	V1	I.+S.	55.3	56.8
MCOf [40]CVPR'18	V1	I.+S.	56.2	57.6
SSNet [48]ICCV'19	V1	I.+S.	57.1	58.6
DSRG [20]CVPR'18	V2	I.+S.	59.0	60.4
SeeNet [19]NeurIPS'18	V1	I.+S.	61.1	60.7
MDC [44]CVPR'18	V1	I.+S.	60.4	60.8
FickleNet [27]CVPR'18	V2	I.+S.	61.2	61.9
OAA [21]ICCV'19	V1	I.+S.	63.1	62.8
ICD [13]CVPR'20	V1	I.+S.	64.0	63.9
Multi-Est. [14]ECCV'20	V1	I.+S.	64.6	64.2
Split. & Merge. [50]ECCV'20	V2	I.+S.	63.7	64.5
SGAN [47]ACCESS'20	V2	I.+S.	64.2	65.0
我们的 EPS	V1	I.+S.	66.6	67.9
	V2	I.+S.	67.0	67.3

表 5. PASCAL VOC 2012 上的分割结果 (mIoU)。所有结果基于 VGG16。所有实验中最优分数以粗体显示。

中的 ResNet101) 的其他现有方法相当甚至更优。我们的方法也显示出相对于现有方法的明显改进。最后,表 6 显示,我们的方法 (在基于 ResNet101 的 DeepLab-V1 和显著性图下) 在 PASCAL VOC 2012 数据集上达到了新的最先进性能 (验证集为 71.0, 测试集为 71.8)。我们强调,现有最先进模型的增益约为 1%,而我们的方法比之前的最佳记录高出 3% 以上。图 5 可视化了我们在 PASCAL VOC 2012 上的

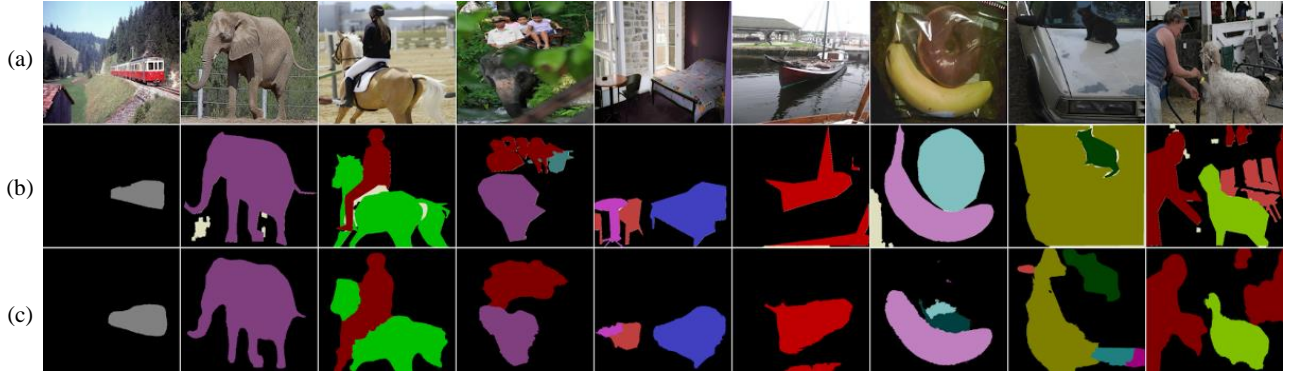


图 6. MS COCO 2014 上分割结果的定性示例。(a) 输入图像, (b) 真实值和 (c) 我们的 EPS。

方法	分割	支持	验证	测试
ICD [13] _{CVPR'20}	V1	I.	64.1	64.3
SC-CAM [5] _{CVPR'20}	V1	I.	66.1	65.9
BES [32] _{ECCV'20}	V2	I.	65.7	66.6
LIID [31] _{TPAMI'20}	V2	I.	66.5	67.5
MCOF [40] _{CVPR'18}	V1	I.+S.	60.3	61.2
SeeNet [19] _{NeurIPS'18}	V1	I.+S.	63.1	62.8
DSRG [20] _{CVPR'18}	V2	I.+S.	61.4	63.2
FickleNet [27] _{CVPR'18}	V2	I.+S.	64.9	65.3
OAA [21] _{ICCV'19}	V1	I.+S.	65.2	66.4
Multi-Est. [14] _{ECCV'19}	V1	I.+S.	67.2	66.7
MCIS [38] _{ECCV'20}	V1	I.+S.	66.2	66.9
SGAN [47] _{ACCESS'20}	V2	I.+S.	67.1	67.2
ICD [13] _{CVPR'20}	V1	I.+S.	67.8	68.0
我们的 EPS	V1	I.+S.	71.0	71.8
	V2	I.+S.	70.9	70.8

表 6. PASCAL VOC 2012 上的分割结果 (mIoU)。所有结果基于 ResNet101。

方法	分割	监督	验证
SEC [25] _{ECCV'16}	V1	I.	22.4
DSRG [20] _{CVPR'18}	V2	I.+S.	26.0
ADL [9] _{TPAMI'20}	V1	I.+S.	30.8
SGAN [47] _{ACCESS'20}	V2	I.+S.	33.6
我们的 EPS	V2	I.+S.	35.7

表 7. MS COCO 2014 上的分割结果 (mIoU)。所有结果基于 VGG16。

分割结果的定性示例。这些结果证实了我们的方法提供了准确的边界, 并成功解决了共现问题。

在表 7 中, 我们进一步在 COCO 2014 数据

集上评估了我们的方法。我们使用基于 VGG16 的 DeepLab-V2 作为分割网络, 与 COCO 数据集的最先进 WSSS 模型 SGAN [47] 进行比较。我们的方法在验证集上达到了 35.7 mIoU, 比 SGAN [47] 高出 1.9%。因此, 我们在 COCO 2014 数据集上达到了新的最先进准确性。在这两个数据集上超越现有最先进方法的出色表现证实了我们方法的有效性; 通过充分利用定位图和显著性图, 它成功地正确捕捉了目标对象的整体, 并弥补了现有模型的不足。图 6 显示了 COCO 2014 数据集上的分割结果的定性示例。当少数对象出现且没有遮挡时, 我们的方法表现良好, 但在处理许多小对象时效果较差。我们的方法的更多示例和失败案例在补充材料中提供。

显著性检测模型的效果。为了研究不同显著性检测模型的效果, 我们采用了三种显著性模型; PFAN [51] (我们的默认模型), DSS [18] 被 OAA [21] 和 ICD [13] 使用, 以及 USPS [34] (i.e., 无监督检测模型)。在基于 Resnet101 的 DeepLab-V1 下的分割结果 (mIoU) 分别为 71.0/71.8 (PFAN), 70.0/70.1 (DSS), 和 68.8/69.9 (USPS) (验证集和测试集)。这些分数支持我们的 EPS 使用任何三种不同的显著性模型仍然比表 6 中的所有其他方法更准确。值得注意的是, 我们的 EPS 使用无监督显著性模型优于所有使用监督显著性模型的现有方法。

6. 结论

我们提出了一种新颖的弱监督分割框架, 即 显式伪像素监督 (EPS)。受定位图和显著性图之间互

补关系的启发, 我们的 EPS 从结合显著性图和定位图的伪像素反馈中学习。由于我们的联合训练方案, 我们成功地补充了两侧的噪声或缺失信息。因此, 我们的 EPS 能够捕捉精确的对象边界, 并丢弃非目标对象的共现像素, 显著提高了伪掩码的质量。广泛的评估和各种案例研究证明了我们 EPS 的有效性和出色的表现, 以及在 PASCAL VOC 2012 和 MS COCO 2014 数据集上 WSSS 的新最先进准确性。

致谢。我们感谢 Duhyeon Bang 和 Junsuk Choe 的反馈。该研究得到了韩国 NRF 基础科学研究计划的支持, 由 MSIP 资助 (NRF-2019R1A2C2006123, 2020R1A4A1016619), 由 MSIT 资助的 IITP 资助 (2020-0-01361, 人工智能研究生院计划 (延世大学)), 以及由韩国政府资助的韩国医疗设备开发基金资助 (项目编号: 202011D06)。

参考文献

- [1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2209–2218, 2019. 2
- [2] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4981–4990, 2018. 2, 6, 7
- [3] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4253–4262, 2020. 2
- [4] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1209–1218, 2018. 5
- [5] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8991–9000, 2020. 2, 8
- [6] Arslan Chaudhry, Puneet K. Dokania, and Philip H. S. Torr. Discovering class-specific pixels for weakly supervised semantic segmentation. In *Proceedings of the British Machine Vision Conference*, 2017. 2
- [7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *International Conference on Learning Representations*, 2015. 5
- [8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017. 5
- [9] Junsuk Choe, Seungho Lee, and Hyunjung Shim. Attention-based dropout layer for weakly supervised single object localization and semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 5, 8
- [10] Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. Evaluating weakly supervised object localization methods right. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3133–3142, 2020. 2
- [11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009. 5
- [12] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015. 5
- [13] Junsong Fan, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4283–4292, 2020. 1, 3, 4, 6, 7, 8
- [14] Junsong Fan, Zhaoxiang Zhang, and Tieniu Tan. Employing multi-estimations for weakly-supervised semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, 2020. 2, 7, 8

- [15] Junsong Fan, Zhaoxiang Zhang, Tieniu Tan, Chunfeng Song, and Jun Xiao. Cian: Cross-image affinity net for weakly supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10762–10769, 2020. 2
- [16] Ruochen Fan, Qibin Hou, Ming-Ming Cheng, Gang Yu, Ralph R Martin, and Shi-Min Hu. Associating inter-image salient instances for weakly supervised semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 367–383, 2018. 2
- [17] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 International Conference on Computer Vision*, pages 991–998. IEEE, 2011. 5
- [18] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3203–3212, 2017. 2, 4, 8
- [19] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In *Advances in Neural Information Processing Systems*, pages 549–559, 2018. 7, 8
- [20] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong 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. 2, 5, 7, 8
- [21] Peng-Tao Jiang, Qibin Hou, Yang Cao, Ming-Ming Cheng, Yunchao Wei, and Hong-Kai Xiong. Integral object mining via online attention accumulation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2070–2079, 2019. 1, 3, 4, 7, 8
- [22] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 876–885, 2017. 1
- [23] Dahun Kim, Donghyeon Cho, Donggeun Yoo, and In So Kweon. Two-phase learning for weakly supervised object localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3534–3543, 2017. 1
- [24] Alexander Kolesnikov and Christoph Lampert. Improving weakly-supervised object localization by micro-annotation. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference*, pages 92.1–92.12. BMVA Press, September 2016. 2
- [25] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 695–711. Springer, 2016. 1, 2, 5, 7, 8
- [26] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in Neural Information Processing Systems*, pages 109–117, 2011. 6, 7
- [27] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5267–5276, 2019. 1, 2, 3, 7, 8
- [28] Kunpeng Li, Ziyang Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9215–9223, 2018. 1, 2, 3, 5, 7
- [29] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3159–3167, 2016. 1
- [30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014. 5
- [31] Yun Liu, Yu-Huan Wu, Pei-Song Wen, Yu-Jun Shi, Yu Qiu, and Ming-Ming Cheng. Leveraging instance-, image-and dataset-level information for weakly supervised instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2, 8

- [32] Chen Liyi, Wu Weiwei, Chenchen Fu, Xiao Han, and Yuntao Zhang. Weakly supervised semantic segmentation with boundary exploration. In *Proceedings of the European Conference on Computer Vision*, 2020. 1, 5, 6, 7, 8
- [33] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014. 5
- [34] Tam Nguyen, Maximilian Dax, Chaithanya Kumar Mummadi, Nhung Ngo, Thi Hoai Phuong Nguyen, Zhongyu Lou, and Thomas Brox. Deepusps: Deep robust unsupervised saliency prediction via self-supervision. In *Advances in Neural Information Processing Systems*, pages 204–214, 2019. 2, 4, 8
- [35] Seong Joon Oh, Rodrigo Benenson, Anna Khoreva, Zeynep Akata, Mario Fritz, and Bernt Schiele. Exploiting saliency for object segmentation from image level labels. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 5038–5047. IEEE, 2017. 2, 5
- [36] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1796–1804, 2015. 1
- [37] Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1713–1721, 2015. 1
- [38] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, 2020. 2, 8
- [39] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 136–145, 2017. 2, 5
- [40] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1354–1362, 2018. 2, 4, 7, 8
- [41] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12275–12284, 2020. 1, 3, 5, 6, 7
- [42] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1568–1576, 2017. 2, 3
- [43] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2314–2320, 2016. 2, 4
- [44] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7268–7277, 2018. 2, 7
- [45] Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90:119–133, 2019. 5
- [46] Huaxin Xiao, Jiashi Feng, Yunchao Wei, Maojun Zhang, and Shuicheng Yan. Deep salient object detection with dense connections and distraction diagnosis. *IEEE Transactions on Multimedia*, 20(12):3239–3251, 2018. 2
- [47] Qi Yao and Xiaojin Gong. Saliency guided self-attention network for weakly and semi-supervised semantic segmentation. *IEEE Access*, 8:14413–14423, 2020. 2, 4, 6, 7, 8
- [48] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, and Lihe Zhang. Joint learning of saliency detection and weakly supervised semantic segmentation. In *Proceedings of the*

IEEE International Conference on Computer Vision, pages 7223–7233, 2019. [2](#), [7](#)

- [49] Bingfeng Zhang, Jimin Xiao, Yunchao Wei, Mingjie Sun, and Kaizhu Huang. Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 12765–12772. AAAI Press, 2020. [2](#)
- [50] Tianyi Zhang, Guosheng Lin, Weide Liu, Jianfei Cai, and Alex Kot. Splitting vs. merging: Mining object regions with discrepancy and intersection loss for weakly supervised semantic segmentation. In Proceedings of the European Conference on Computer Vision, 2020. [2](#), [7](#)
- [51] Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3085–3094, 2019. [1](#), [2](#), [3](#), [4](#), [5](#), [8](#)
- [52] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2921–2929, 2016. [1](#), [5](#), [6](#), [7](#)