铁路不是火车：在弱监督语义分割中使用显著性图作为伪像素级监督信号

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

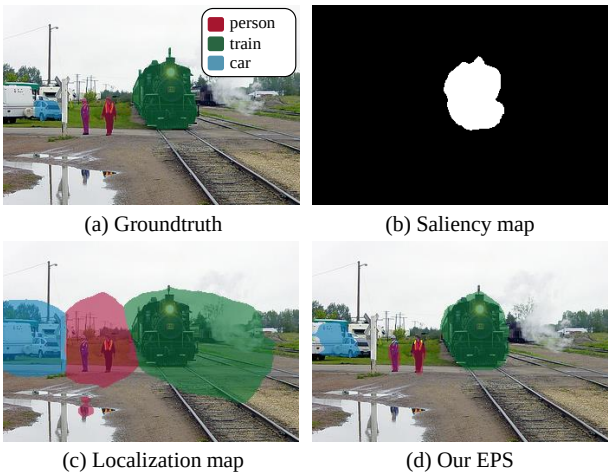
当前有关弱监督语义分割（WSSS）的使用图像级弱监督信号的研究有以下一些限制：松弛的目标覆盖，以及来自于非目标物体的重复像素。为了解决这些问题，我们提出了一种新颖的框架，并给其命名为详确伪像素级监督（EPS）。我们的模型主要是通过结合两种像素级监督信号的反馈来实现分割；像素级标签通过定位图来提供物体识别信息，还有从现成的显著性检测模型中提取到的显著性图提供的丰富的边界信息。我们设计了一种联合训练方案来充分利用这两种信息间的互补关系。我们的方法能够获得准确的物体边界并且丢弃掉重复像素，进而可以显著地提高伪标签的质量。实验结果显示我们提出的方法明显地比其它现成的方法在弱监督语义分割的主要比赛中表现要好。其中，在 PASCAL VOC 2012 和 MS COCO 2014 数据集上取得了最好的成绩。代码可以在以下链接中找到 https://github.com/halbielee/EPS.

1. Introduction

弱监督语义分割（WSSS）使用弱监督信号 (e.g.，图像级标签 [36, 37]，涂鸦 [29]，或者范围框 [22]) 并且致力于达到与需要像素级标签的全监督模型达到可以相比的性能。大部分已有的研究采用图像级标签来作为分割模型的弱监督信号。WSSS 的整体流程大体分为两个阶段。首先，使用一个图像的分类器来为目标物体产生伪标签。然后，使用这些为标签作为监督信号来训练分割模型。最流行的产生伪标签的技巧是类激活映射（CAM）[52]，CAM 可以提供与它们图像级标签相同的物体定位图。由于全监督语义分割 (i.e., 像素级标签) 和弱监督语义分割 (i.e., 图像级标签) 监督信号之间的差距，WSSS 有以下几种关键挑战：1) 定位图只捕捉目标物体的一小部分 [52], 2) 存在物体边界不匹配 [23], 和 3) 很难区分开来自目标物体的重复像素 (e.g., 铁路和火车) [25]。

为了解决这些问题，现有的研究可以被分为三类。第一种方法通过扩展物体的覆盖范围来捕捉目标的全部，例如擦除像素 [9, 23, 28]，组装得分图 [21, 27]，或者使用自监督信号 [41]。但是，由于缺少对物体形状的指导，这些方法都不能确定准确的目标物体边界。第二种方法则着重于提升伪标签的物体边界 [13, 32]。由于它们可以有效地学习目标边界，因此自然而然的，它们能扩展伪标签到边界。然而，它们还是无法从目标物体中区分出来自非目标物体的偶然像素。这是因为前景和背景之间的那种很强的关联 (i.e., 重复) 是几乎无法从归纳偏向中区分出来的 (i.e., 观察到目标物体或者与之相对的偶然像素的频率)，这正如 [10]中表述的那样。最后，第三种办法志于通过使用额外的真实标签 [24]，或者显著性图 [35, 47]来缓解重复问题。但是，[24, 28] 需要很强的像素级标注，这与弱监督学习的宗旨相

*表示相等的贡献。
†Hyunjung Shim 是通讯作者。



(a) Groundtruth (b) Saliency map

(c) Localization map (d) Our EPS

图 1. 在 WSSS 中使用定位图和显著性图的启发样例。(a) 真是标签，(b) 显著性图提供于 PFAN [51]，(c) 定位图提供于 CAM [52] 和 (d) EPS 使用定位图和显著性图来训练分类器。可以看出显著性图不关注 人和 车，然而我们的方法可以正确地存储它们，从而定位图能关注到它们。

悖。[35] 对显著性图的错误很敏感。此外，[47]不能覆盖目标的全部并且存在边界不匹配问题。

在这篇文章中，我们的目标是解决 WSSS 存在的三种主要挑战，通过使用定位图 (i.e., 使用图像级标签训练的分类器生成的 CAM) 和显著性图 (i.e., 现成的显著性检测模型的输出 [18, 34, 51])。我们关注这两种图之间的互补关系。就如同图 1展示的那样，定位图可以区分不同的物体但是不能有效地分离它们的边界。相对的，显著性图提供了丰富的边界信息，但是忽略了物体的区分。因此就此点来说，我们认为这种使用这两部分互补信息的方法可以突破 WSSS 的性能瓶颈。

最后，我们提出了一种针对 WSSS 的新颖的框架，并命名为 *详确伪像素级监督* (EPS). 为了充分利用显著性图 (i.e., 前景和背景), 我们设计了一种可以预测 $C+1$ 个类的分类器, 由 C 个目标类和背景类组成。我们利用 C 个定位图和背景的定位图来组合一个显著性图。然后，显著性损失被定义为真实显著性图和我们的组合显著性图之间的像素级差异。通过引进显著性损失，我们的模型可以被所有类的像素级反馈监督。我们也使用了多类的分类损失来预测图像级标签。因此，我们训练我们的分类器通过优化显著性损失和多类的分类损失，并且协同增强对背景像素和前景像素的预测– 我们发现我们的策略不仅可以提升显著性图 (Section 3.3 和 Figure 3) 还可以提升伪标签 (Section 5.1 和图 4).

我们认为，由于显著性损失通过伪像素级反馈惩罚边界不匹配，其可以促使我们的模型去学习目标的准确边界。附加的，我们能够通过扩展图片到边界捕捉到整个物体。由于显著性图帮助分离前景 (e.g., 火车) 和背景, 我们的方法可以分配重复像素到 (e.g., 铁路) 背景类. 实验结果表明我们的 EPS 得到值得称赞的分割性能，其在 PASCAL VOC 2012 和 MS COCO 2014 数据集都取得了第一的好成绩。

2. Related Work

弱监督语义分割。 WSSS 通用的流程是先通过一个分类网络产生伪标签然后使用这些伪标签作为监督信号来训练分割网络。由于图像级标签缺乏边界信息，许多现存的方法都存在不准确的伪标签的问题。为了解决这些问题，跨图片的亲和矩阵 [15]，知识图 [31] 和对比优化 [38, 50] 被使用来提升伪标签。[5] 提出了一种自监督任务探索子类来推动分类器产生更好的 CAM。[1, 2] 通过计算像素级的相似矩阵来隐式地挖掘边界信息。[49] 则关注于产生可靠的像素级标注并且设计了端到端的网络来产生分割图。[20, 25] 使用了一种边界损失来训练分割模型。最近，[3] 使用了一种带有自监督训练策略的单一分割模型。[14] 则使用多种不完整的伪标签来强调分割网络的鲁棒性。

显著性指引的语义分割。显著性检测（SD）通过额外的带有像素级标注 [18, 46, 51]或者图像级标注 [39]的显著性数据集来产生显著性图，然后对一张图片中的前景和背景进行区分。许多 WSSS 方法 [15, 20, 27, 28, 42, 44] 采用显著性图来作为伪标签中背景的指引。[43] 利用显著性图作为单目标图片的全监督信号。[16] 使用实例级的显著性图来学习目标间的相似图。[6, 40, 47] 结合了显著性图和针对类的注意力机制来生成伪标签。[48] 使用了单一网络同时提高了 WSSS 和 SD 的性能。我们的 EPS 可以被分类在显著性指引的方法中，但是又与其他的方法有着明显的不同。具体的原因可以从以下几

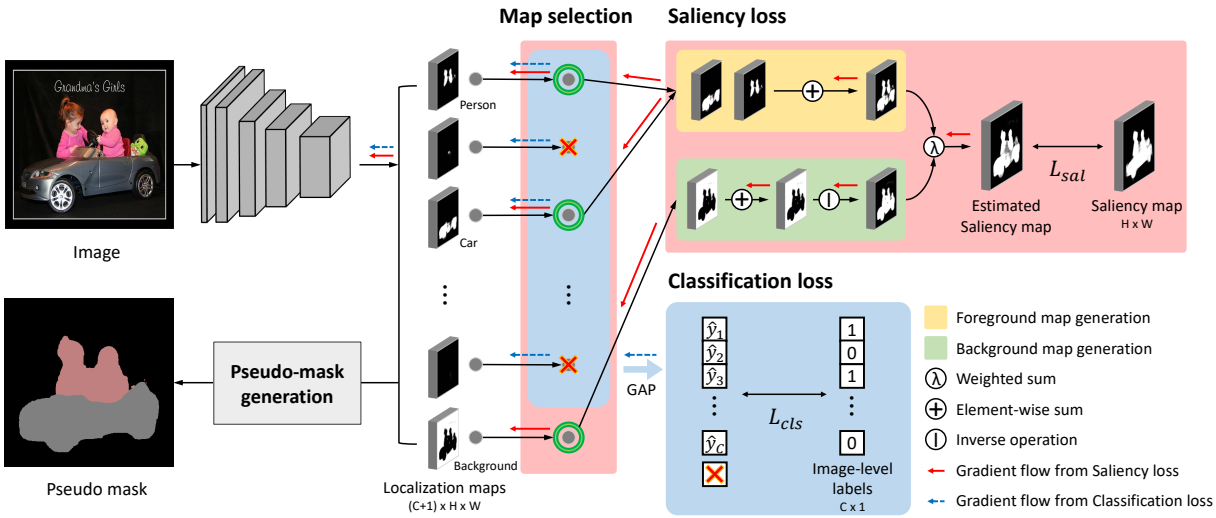


图 2. EPS 的整体框架。$C+1$ 个定位图由主干网络生成。真实显著性图由现成的显著性检测模型生成。一些定位图被有选择地用于生成估计显著性图 (Section 3.2)。整体框架使用显著性损失和分类损失联合训练 (Section 3.3)。

个方面来说。大部分已有的方法采用显著性图作为伪标签的一部分或者隐式地作为精炼分类器中间特征的指引。与之不同的是，我们的方法使用显著性图作为像素级的对定位图的反馈。尽管 [48] 在利用两种互补信息这个方面来说是与我们的工作最相像的工作，但是它既没有解决重复的问题也没有对那些有噪声的显著性图进行处理。

3. Proposed Method

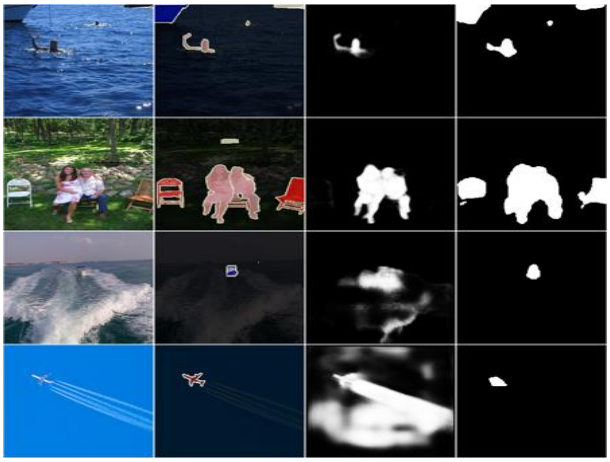
在这个部分，我们针对弱监督语义分割 (WSSS) 提出了一种新的框架，并命名为 *详确伪像素级监督* (EPS)。考虑到在 WSSS 中的两个阶段，我们的第一个阶段也是先产生伪标签，然后第二个阶段去训练分割模型。在这上面来说，我们的主要贡献是产生准确的伪标签。按照 WSSS 的惯例 [13, 21, 27, 28, 41, 42]，我们使用生成的伪标签作为训练分割模型的监督信号。

3.1. Motivation

我们的核心思路是充分利用两种互补信息 i.e.，定位图中的目标识别和显著性图的边界信息。为了这个目的，我们使用显著性图来作为前景和背景定位图的伪像素级反馈。我们设计带有额外背景类的分类器，最终预测总共 $C+1$ 个类，正如图 2所展示的那样. 使用这个分类器，我们可以学到 $C+1$ 个定位图，i.e., C 个对目标标签的定位图和一个背景的定位图。

紧接着，我们阐释 EPS 是如何解决 WSSS 中边界不匹配和重复像素的问题的。为了解决边界不匹配问题，我们用 C 个定位图估算前景图，然后把它和前景的显著性图匹配。用这种方法，定位图能得到伪像素级的来自显著性图的反馈，从而增强目标的边缘。为了去除来自非目标物体的重复像素，我们也将背景的定位图和相应的显著性图匹配。因为背景的定位图也可以得到伪像素级的来自显著性图的反馈，重复像素可以被成功地分配给背景；来自非目标的重复像素最大程度地与背景重叠。这就解释了为什么我们的方法可以从目标物体分离出重复像素。

最终，EPS 的目标函数最终有两部分组成：来自显著性图的显著性损失 $\mathcal{L}_{sal}$ (在图 2中用红色线或箭头标出)；和来自图像级标签的多类分类损失 $\mathcal{L}_{cls}$ (在图 2用蓝色线或箭头标出)。通过联合训练这两个目标损失，我们可以通过两者间的互补信息协同增强定位图和显著性图– 我们观察到使用我们的联合训练策略，两者的噪声和缺损的信息都被补充，就如同图 3展示的。举例来说，原始的从现成模型中得到的显著性图 [18, 34, 51]有缺失的和噪声信息。而我们的结果却能够保存缺失的目标 (e.g., 船和椅子)



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

图 3. 在 PASCAL VOC 2012 数据集上的定量的分割结果 (a) 输入图片，(b) 真实标签，(c) [51]中的显著性图，和 (d) 我们的估算显著性图

并且移除噪声 (e.g., 水的波纹和凝迹)，明显胜过原显著性图。因此从结果来看，EPS 可以捕捉更准确的物体边界并且从目标物体分离出重复像素。这些优势形成了显著的性能提升；表 6 揭示了 EPS 显著地比其他模型在分割准确率上有 3.8–10.6% 的提升。

3.2. Explicit Pseudo-pixel Supervision

我们解释了怎样使用显著性图作伪监督。显著性图的主要优势就在于提供了物体轮廓，从而能更好地揭示物体边界。为了更好地利用这个优势，我们将显著性图与两个实例进行比较：前景和背景。为了让类级的定位图与显著性图具有可比性，我们融合目标的所有定位图并最终生成一个前景图，$\mathrm{M}_{fg} \in \mathbb{R}^{H\times W}$。我们也使用背景图的倒置来表示前景。其中背景图是背景标签 $\mathrm{M}_{bg} \in \mathbb{R}^{H\times W}$ 对应的定位图。(之后，我们会详细解释如何精炼前景图来处理有噪声的显著性图。)

具体来说，我们估算显著性图 $\widehat{\mathrm{M}}_s$ 通过 M_{fg} 和 M_{bg} ，公式如下：

$$\widehat{\mathrm{M}}_s = \lambda \mathrm{M}_{fg} + (1-\lambda)(1-\mathrm{M}_{bg}), \tag{1}$$

$\lambda \in [0,1]$ 是一个超参数来调整前景之和和背景反转的权重 (默认的，在实验中，我们设置 λ 为 0.5 并且关于 λ 的消融实验可以在补充材料里找到。) 紧接着，我们定义显著性损失 $\mathcal{L}_{sal}$ 为估算的显著性图和真实显著性图之间像素级差别的和 ($\mathcal{L}_{sal}$ 的公式定义在 Section 3.3给出)

值得一提的是，在弱监督语义分割中使用预训练模型是被鼓励的，因此在 WSSS 中使用显著性图是被广泛接受为一种寻常策略的。尽管这很流行，采用全监督显著性检测模型仍是具有争议的，因为它们会使用来自不同数据集的像素级标注。在这篇文章中，我们观查了不同显著性图检测模型的效果；1) 无监督和 2) 全监督显著性图检测模型 (见 Section 5.3)，并且实验也证明我们的方法不论使用哪一种都明显优于其他全监督显著性模型 [13, 21, 40, 43, 47]。鉴于现有模型没有充分利用显著性图，我们的方法使用显著性图作为伪像素级监督信号，并将其作为边界和重复像素的指引。

映射选择解决显著性偏差。在前面提到，我们假设前景图可以是所以目标的定位图融合；背景图可以是背景的定位图。但是，如此天真的设计方案可能使我们的显著性图与通过现成模型计算出的显著性图无法相比。比如说，显著性图来自 [51]经常忽略一些目标作为显著性目标 (e.g., 图 1中靠近火车的很小的人). 这个系统误差是无法避免的，因为显著性模型学习不同数据集的统计特征。

为了解决这个系统误差，我们发明了一种有效的策略，我们使用显著性图和定位图的重叠率。具体来说，第 i 个定位图 M_i 被分配到前景，当 M_i 其与显著性图重叠超过 $\tau\%$，否则分配到背景。其计算公式如下：

$$\begin{aligned} \mathrm{M}_{fg} &= \sum_{i=1}^{C} y_i \cdot \mathrm{M}_i \cdot \mathbb{1}[\mathcal{O}(\mathrm{M}_i, \mathrm{M}_s) > \tau], \\ \mathrm{M}_{bg} &= \sum_{i=1}^{C} y_i \cdot \mathrm{M}_i \cdot \mathbb{1}[\mathcal{O}(\mathrm{M}_i, \mathrm{M}_s) \le \tau] + \mathrm{M}_{C+1}, \end{aligned} \tag{2}$$

$y \in \mathbb{R}^C$ 是二值的图像标签，$\mathcal{O}(\mathrm{M}_i, \mathrm{M}_s)$ 是计算 M_i 和 M_s 之间重叠率的函数。为了这个目的，我们首先二值化定位图和显著性图如下：对于像素 p $\mathrm{B}_k(p) = 1$ 当 $\mathrm{M}_k(p) > 0.5$; $\mathrm{B}_k(p) = 0$，其余情况。B_i 和 B_s 是和 M_i 与 M_s 相对应的二值化后的图。紧接着，我们计算 M_i 和 M_s 之间的重叠率，i.e., $\mathcal{O}(\mathrm{M}_i, \mathrm{M}_s) = |\mathrm{B}_i \cap \mathrm{B}_s| / |\mathrm{B}_i|$。我们设置 $\tau = 0.4$ 不论使用什么数据集和主干。在补充材料中，我们验证的我们的结果

对 τ 的选取是鲁棒的。(i.e., τ 在 [0.3, 0.5] 之间表现出相似的性能)。

相比单一的定位图作为背景，我们结合背景的定位图和没有被选为前景的定位图作为背景。尽管很简单，但是我们能够避免显著性图的错误，并且有效地训练出被显著性图忽略的一些目标。(在表 3，我们展示了提出的策略在解决显著性图错误的有效性。)

3.3. Joint Training Procedure

使用显著性图和图像级标签，EPS 总的损失函数由两部分组成，显著性损失 $\mathcal{L}_{sal}$ 和分类损失 $\mathcal{L}_{cls}$。首先，显著性损失 $\mathcal{L}_{sal}$ 通过计算真实显著性图 M_s 和估算显著性图 $\widehat{\mathrm{M}}_s$ 之间的平均像素级距离组成。

$$\mathcal{L}_{sal} = \frac{1}{H \cdot W} ||\mathrm{M}_s - \widehat{\mathrm{M}}_s||^2, \tag{3}$$

M_s 通过现成的显著性检测模型– PFAN [51] 训练在 DUTS [39]数据集上得到。值得注意的是，我们的方法比之前的方法要好不论采用何种显著性检测模型。

接下来，分类损失通过像素级标签 y 和它的预测 $\hat{y} \in \mathbb{R}^C$ 之间的多标签 softmax 损失计算，其中预测是通过定位图在每一个目标类上的全局平均池化得到的。

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

$\sigma(\cdot)$ 是 sigmoid 函数。最终，整体训练损失是多标签分类损失和显著性损失之和，i.e., $\mathcal{L}_{total} = \mathcal{L}_{cls} + \mathcal{L}_{sal}$.

正如图 2, $\mathcal{L}_{sal}$ 参与更新 $C+1$ 个类的参数，包含目标物体和背景。与此同时，$\mathcal{L}_{cls}$ 仅仅评价了 C 类的预测结果，不包含背景类– $\mathcal{L}_{cls}$ 的梯度不会影像背景类的参数更新。但是，背景类可以被 $\mathcal{L}_{cls}$ 隐式的影响，因为显著性损失监督分类器的训练。

4. Experimental Setup

数据集。我们在两个流行的基准数据集上做了实验研究, PASCAL VOC 2012 [12] 和 MS COCO 2014 [30]。PASCAL VOC 2012 由 21 个类组成 (i.e., 20 目标还有背景)，其中分别有 1464, 1449, 1456 张图片作为训练集，测试集，验证集。就像之前语义分割做的那样，我们使用增广后的训练集，共有 10582 张图片作为训练集 [17]。然后，COCO 2014 包含 81 个类，其在一个类是背景，其中由 82081 张图片用来训练，40137 张用来验证，没有目标类的图片被剔除，就想 [9]中做的一样。因为对其中一些物体的真实的分割标签有互相的重叠，我们采用 COCO-Stuff [4]中的真实分割标签，其解决了同一个 COCO dataset 的重复问题

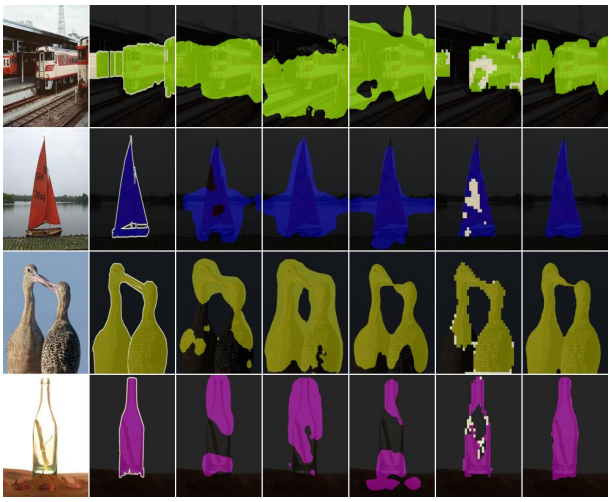
评价标准。我们在 PASCAL VOC 2012 的验证集和测试集以及 COCO 2014 的验证集上验证了我们的方法。在 PASCAL VOC 2012 上的验证结果是从官方 PASCAL VOC 2012 的验证服务商得到的。此外，我们采用了平均交并比来策略分割模型的准确性。

实现细节。我们选择了 ResNet38 [45] ，输出步长为 8，作为我们方法的主干网络。所有的主干网络均在 ImageNet [11]上进行了预训练。我们使用了 SGD 优化器，并设批大小为 8。我们的方法设学习率为 0.01 (最后的卷积层设为 0.1) 并且模型最终迭代 20k 次。就数据增强来说，我们使用了随机拉伸，随机翻转，以及随机裁剪到 448×448。分割模型我们选取了 DeepLab-LargeFOV (V1) [7] 和 DeepLab-ASPP (V2) [8], 还有 VGG16 和 ResNet101 的主干网络。具体来说，我们使用了四种分割网络: VGG16 为基础的 DeepLab-V1 和 DeepLab-V2, ResNet101 为基础的 DeepLab-V1 和 DeepLab-V2。更多细节请见补充材料。

5. Experimental Results

5.1. Handling Boundary and Co-occurrence

边界不匹配问题。为了验证伪标签的边界性能，我们比较了边界质量同最好的一些方法 [32, 41, 52]。我们选取了 SBD [17]，其在 PASCAL VOC 2011 提供了边界标注和边界的基准。就如同 [32]所做，边界质量被预测结果和真是标签之间的回归率，准确率和 F1-score 来衡量。表 1 揭示了我们的方法在提出的三个指标上都强过其他方法。定量结果，展示在图 4 则表明我们的方法与其他方法相比能捕捉到更准确的边界信息。



(a) (b) (c) (d) (e) (f) (g)

图 4. PASCAL VOC 2012 上的定量样例结果 (a) 输入图片，(b) 真实标签，(c) CAM,，(d) SEAM，(e) ICD，(f) SGAN 和 (g) our EPS。

Method	Recall (%)	Precision (%)	F1-score (%)
CAM [52]$_{\text{CVPR'16}}$	22.3	35.8	27.5
SEAM [41]$_{\text{CVPR'20}}$	40.2	45.0	42.5
BES [32]$_{\text{ECCV'20}}$	45.5	46.4	45.9
Our EPS	60.0	73.1	65.9

表 1. 在 SBD 训练验证集上测试的边界准确性。值得注意的是，BES 的结果是由提出在 [32]的边界预测网络测量的。

重复像素问题。就像在一些研究中 [20, 25, 28, 35]讨论得那样，我们观察到在 PASCAL VOC 2012 中，一些背景类频繁地和目标物体一起出现。我们使用 PASCAL-CONTEXT 数据集 [33]定量地分析了重复目标的频率，在 PASCAL-CONTEXT 数据集中提供了一张图中所有的像素级标注，(e.g., 水 and 铁路)。我们选择了三种重复出现的序列；船和 水，火车和 铁路，还有 火车和 站台。我们比较了目标类还有它相应的重复类之间的目标类 IoU 和 混淆矩阵。混淆率反映了重复类被正确预测为目标类的程度。混淆率 $m_{k,c}$ 的计算如 $m_{k,c} = FP_{k,c}/TP_c$，其中 $FP_{k,c}$ 是重复类 k 被错误分为目标类 c 的像素数，TP_c 是目标类 c 的真阳性像素。更多重复问题的细节分析呈现在补充材料中。

表 2 展示了 EPS 一致地表现出来比其他方法都要好低的混淆率。SGAN [47] 有和我们相近的混淆率，但是我们的方法能够在 IoU 准确率上更好低捕捉到目标。这是因为 SEAM [41] 尝试应用自监督训练来覆盖目标物体的全部，这会使得其更容易被目标类的重复像素迷惑。与此同时，CAM 只关注目标物体之间最具有区别性的区域，并不会覆盖那些不那么具有区别性的区域。e.g., 重复类。我们也能看到这些现象在图 4中。

Method	boat w/ water	train w/ railroad	train w/ platform
CAM [52]$_{\text{CVPR'16}}$	0.74 (33.1)	0.11 (52.9)	0.09 (49.6)
SEAM [41]$_{\text{CVPR'20}}$	1.13 (30.7)	0.24 (48.6)	0.20 (45.5)
ICD [13]$_{\text{CVPR'20}}$	0.47 (41.4)	0.11 (56.7)	0.09 (49.2)
SGAN [47]$_{\text{ACCESS'20}}$	0.10 (42.3)	0.02 (48.8)	0.01 (36.3)
Our EPS	0.10 (55.0)	0.02 (78.1)	0.01 (73.0)

表 2. 在解决重复中与代表性的其他已有方法的比较。每一个条目是$m_{k,c}$ 为 blue (越低越好) 和 IoU 在括号中 (越高越好)。

	Baseline	Naïve	Pre-defined	Our adaptive
mIoU	66.1	66.5	67.9	69.4

表 3. 映射选择策略的影响。在 PASCAL VOC 2012 训练集上使用不同映射选择策略的伪标签准确率。

5.2. Effect of Map Selection Strategies

我们评估了我们的映射选择策略对消除显著性图错误的贡献。我们比较了不同的映射选择策略和没有使用任何映射选择策略的基线。对于单纯的一般策略，前景图是所有目标定位图的融合；背景图则等于背景类的定位图。(i.e., 天真策略)。紧接着是根据一些例外延伸的一般策略。一些预定义类的定位图，(e.g., 沙发，长椅，和 饭桌) 被分配给背景图，(i.e., 预定义类策略)。最后，提出的选择策略使用重定位图和显著性图之间的重复覆盖率，正如 Section 3.2 (i.e., 我们的自适应策略)。

表 3揭示了我们适应性的策略能够有效地解决显著性图的系统不一致问题。那些单纯的策略在生成估计显著性图时没有展示出任何对一致性的考虑。就这方面来说，伪标签的效果会发生衰退，特别是

图 5. 在 PASCAL VOC 2012 数据集上的定量分割结果样例 (a) 输入图片，(b) 真实标签，和 (c) 我们的 EPS。

Method	w/o refinement	w/ CRF [26]	w/ AffinityNet [2]
CAM [52]$_{\text{CVPR'16}}$	48.0	-	58.1
SEAM [41]$_{\text{CVPR'20}}$	55.4	56.8	63.6
ICD [32]$_{\text{CVPR'20}}$*	59.9	62.2	-
SGAN [47]$_{\text{ACCESS'20}}$*	62.8	-	-
Our EPS	69.4	71.4	71.6

表 4. 在 PASCAL VOC 2012 上测试的伪标签的准确率 (mIoU) 。* 表示低置信度的像素被忽略，其余的全部像素都参与测试。

在 沙发，长椅或者 饭桌类上。使用预定义的类表现出忽略显著性图中缺失的类会使得不一致得到缓解。但是，它需要观察者进行手动筛选，这不现实而且不能对每张图做出最优的决策。然而，我们的适应性的策略可以自动解决一致性问题并且对给定的显著性图做出更有效的决策。

5.3. Comparison with state-of-the-arts

伪标签的准确率。我们采用了一种多尺度的推断方案，通过不同尺度增强图片的预测结果，这是常见的方式在 [2, 41]。然后，我们在训练集通过比较我们的 EPS 和基线 CAM [52] 和三种最优的模型 i.e., SEAM [41], ICD [13], and SGAN [47]评估伪标签的准确率。特别说明，WSSS 中在训练集上估计伪标签的正确率是非常常见的方法，这些伪标签也会在之后用于训练分割模型。表 4 总结了伪标签的准确率并且证实了我们的方法能够明显的超出其他的现有方法 (i.e., 7–21% 的差距)。图 4 定量地可视化了

Method	Seg.	Sup.	val	test
SEC [25]$_{\text{ECCV'16}}$	V1	I.	50.7	51.7
AffinityNet [2]$_{\text{CVPR'18}}$	V1	I.	58.4	60.5
ICD [13]$_{\text{CVPR'20}}$	V1	I.	61.2	60.9
BES [32]$_{\text{ECCV'20}}$	V1	I.	60.1	61.1
GAIN [28]$_{\text{CVPR'18}}$	V1	I.+S.	55.3	56.8
MCOF [40]$_{\text{CVPR'18}}$	V1	I.+S.	56.2	57.6
SSNet [48]$_{\text{ICCV'19}}$	V1	I.+S.	57.1	58.6
DSRG [20]$_{\text{CVPR'18}}$	V2	I.+S.	59.0	60.4
SeeNet [19]$_{\text{NeurIPS'18}}$	V1	I.+S.	61.1	60.7
MDC [44]$_{\text{CVPR'18}}$	V1	I.+S.	60.4	60.8
FickleNet [27]$_{\text{CVPR'18}}$	V2	I.+S.	61.2	61.9
OAA [21]$_{\text{ICCV'19}}$	V1	I.+S.	63.1	62.8
ICD [13]$_{\text{CVPR'20}}$	V1	I.+S.	64.0	63.9
Multi-Est. [14]$_{\text{ECCV'20}}$	V1	I.+S.	64.6	64.2
Split. & Merge. [50]$_{\text{ECCV'20}}$	V2	I.+S.	63.7	64.5
SGAN [47]$_{\text{ACCESS'20}}$	V2	I.+S.	64.2	65.0
Our EPS	V1	I.+S.	66.6	67.9
	V2	I.+S.	67.0	67.3

表 5. 在 PASCAL VOC 2012 数据集上的分割结果 (mIoU) on PASCAL VOC 2012。All results are based on VGG16. 所有实验中的最好结果进行黑体表示。

样本的伪标签，证明了我们的方法可以明显提升目标边界并且定量地在伪标签的准确性上超出最好的方法。我们的方法能准确地获得目标边界 (第二行) 并且由此自然地覆盖目标的全部 (第三行)，此外还能缓解重复像素问题 (第一行)。更多例子还有我们方法的失败方案都在补充材料中提供。

分割结果的准确性。之前的方法 [2, 13, 41] 生成伪

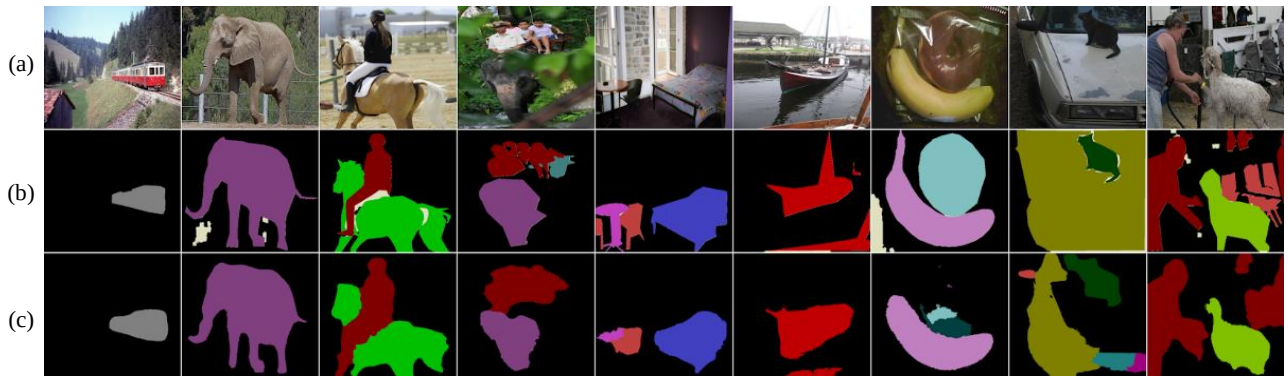


图 6. 在 MS COCO 2014 数据集上的定量分割结果样例 (a) 输入图片，(b) 真实标签，和 (c) 我们的 EPS。

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

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

Method	Seg.	Sup.	val
SEC [25]$_{\text{ECCV'16}}$	V1	I.	22.4
DSRG [20]$_{\text{CVPR'18}}$	V2	I.+S.	26.0
ADL [9]$_{\text{TPAMI'20}}$	V1	I.+S.	30.8
SGAN [47]$_{\text{ACESS'20}}$	V2	I.+S.	33.6
Our EPS	V2	I.+S.	35.7

表 7. 在 MS COCO 2014 数据集上的分割结果 (mIoU)。所有结果基于 VGG 16。

标签并且用 CRF 算法 [26] 或者用亲和网络 [2]来精炼它们。然而，正如表 4所示，我们生成的伪标签已经足够准确，因此我们没有添加任何伪标签的精炼策略来训练分割网络。我们在 PASCAL VOC 2012 数据集上扩展地测试并准确地比较了我们的方法和其他四种分割网络。

我们的方法不论选取何种分割网络都明显优于其他网络。表 5 展示了使用相同的 VGG 16 主干，我们的方法比其他的方法更加准确。此外，我们使用 VGG 16 的结果能够取得和使用更强大的主干网络，相匹敌甚至更好的性能 (i.e., 在表 6中的 ResNet101)。我们的方法也清晰地表明了相比于其他的进步。最后，表 6 阐明了我们的方法 (基于使用了 Deeplab-V1 的 ResNet101) 在 PASCAL VOC 2012 数据集上取得了新的最好的结果 (验证集 71.0，测试集 71.8)。我们强调，相比其他最好的方法取得的性能提升有接近 1%。此外，我们的方法达到了相比之前最好的记录超出 3% 的提升。图 5 定性地可视化了在 PASCAL VOC 2012 上的分割结果样例。这些结果证实了我们的方法提供了更加确切的边界，并且成功解决了重复问题。

在表 7中，我们在 COCO 2014 数据集上进一步评估了我们的方法。我们使用基于 DeepLab-V2 的 VGG 16 作为分割网络来与 SGAN [47]，在 COCO 数据集上最好的 WSSS 模型，进行比较。我们的方法获得了验证集上 35.7 的 mIoU 并且它超出 SGAN [47] 了 1.9%。就结果来说，我们获得了 COCO 2014 数据集上的最高准确率。这些在两个数据集上相比其他最好的方法取得的杰出结果证实了我们方法的有效；通过充分利用定位图和显著性图，成功正确地捕捉到了目标的整体并且弥补了现有模型的

缺陷。图 6 展示了在 COCO 2014 数据集上的定性分割结果。我们的方法在一些没有闭塞的目标出现时表现很好，但是在处理许多小目标时不那么有效。更多例子和失败案例在补充材料中提供

显著性检测模型的效果。为了研究不同的显著性检测模型的效果，我们选用了三个显著性模型；PFAN [51] (我们的默认模型)，DSS [18] 使用于 OAA [21] 和 ICD [13]，还有 USPS [34] (i.e., 无监督检测模型)。使用基于 DeepLab-V 的 Resnet101 的分割结果 (mIoU) 为 71.0/71.8 相比于 PFAN，70.0/70.1 相较于 DSS，68.8/69.9 相较于 USPS (验证集和测试集)。这些结果支持 EPS 不论是用何种有效的显著性检测模型都可以取得比其他方法更好分割结果的说法，具体见表 6。值得一提的是，EPS 使用无监督显著性检测模型也由于那些已有的使用有监督显著性模型的方法。

6. Conclusion

我们提出了一种新的弱监督语义分割模型，命名为 *详确伪像素级监督* (EPS)。受启发于定位图和显著性图之间的互补关系，EPS 学习结合定位图和显著性图的像素级反馈。得益于我们的联合训练策略，我们成功补足了两者的噪声或者缺失信息。就结果来说，EPS 能获得准确的目标边界并且丢弃来自非目标物体的重复像素，显著地提升了伪标签的质量。大量的实验和各种研究揭示了 EPS 的有效和杰出性能，以及就 WSSS 来说在 PASCAL VOC 2012 和 MS COCO 2014 数据集上取得了最佳。

致谢。我们感谢 Duhyeon Bang 和 Junsuk Choe 的反馈。这项研究受支持于 the Basic Science Research Program 通过 the MSIP (NRF-2019R1A2C2006123, 2020R1A4A1016619) 资助的 the NRF Korea，IITP 承诺受资助于 the MSIT (2020-0-01361, Artificial Intelligence Graduate School Program (YONSEI UNIVERSITY))，和承诺受资助于 the Korean government (Project Number: 202011D06) 的 the Korea Medical Device Development Fund 。

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, 7, 8

[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. 1

[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, 9

[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, 3, 9

[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, 6, 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, 9

[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. 1

[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, 6, 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. 7, 8

[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, Ziyan 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, 6, 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. 6

[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, 3, 9

[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. 1, 2, 6

[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. 1, 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, 3, 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. 2, 3, 4, 5, 9

[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, 2, 5, 6, 7