《Railroad is not a Train: Saliency as Pseudo-pixel Supervision for Weakly Supervised Semantic Segmentation》[1]复现和优化

刘昱志

摘要

现有的使用图像级弱监督的弱监督语义分割 (WSSS) 研究有几个局限性：对象覆盖范围稀疏、对象边界不准确以及来自非目标对象的共现像素。为了克服这些挑战，目前的解决框架名为显式伪像素监督（EPS），它通过结合两个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，现成显着性检测模型的显着性图提供丰富的边界。该框架提出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。在本次工作中，利用语义亲和网络对生产的分割掩码进行进一步优化，提高分割的性能。实验结果表明，加入了语义亲和网络后对分割性能提升了 1.2-2.1 个百分点。

关键词：弱监督；图像处理；语义分割

1 引言

深度神经网络 (DNN) 的最新发展推动了语义分割的显著改进。但语义分割主要障碍之一是缺乏训练数据，由于像素级分割标签的注释成本高得令人望而却步，现有数据集通常缺乏注释示例和类别多样性, 这使得传统方法仅限于数据集中预定义的小范围对象类别。

弱监督语义分割 (WSSS) 利用弱监督（例如，图像级标签、草图或边界框），旨在实现与全监督模型的竞争性能。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），它提供与其图像级标签相对应的对象定位图。由于完全（即像素级注释）和弱（即图像级标签）监督语义分割之间的监督差距目前 WSSS 面临如下挑战：1）定位图仅捕获一小部分目标对象，2) 它受到对象的边界不匹配的影响，以及 3) 它几乎无法将同时出现的像素与目标对象（例如，火车上的铁路）分开。

在论文中，作者提出了一个 WSSS 框架, 目标是通过充分利用定位图（即来自使用图像级标签训练的图像分类器的 CAM）和显着性图（Saliency Map）来克服 WSSS 的三个挑战

2 相关工作

2.1 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性，许多现有方法都存在不准确的伪掩码问题。为了解决这个问题，使用跨图像亲和力[2]、知识图[3] 、对比优化[4] 和协同学习[5]来提高伪掩码的质量。还有使用基于因果干预的上下文调整模型[6]来增强相同像素之间的关联。

此外，许多技术来优化 CAM 来提升分割质量。一些代表性的方法利用图像级隐藏和擦除操作来驱动分类器关注对象的不同部分。[7-9]。有的工作将 CAM 激活区域视为初始“种子”，并逐渐增长种子区域，直到覆盖完整的对象[10-11] 还有一些工作使用扩张卷积[12]、多尺度特征融合来增强 CAM[13]。

2.2 基于 Saliency 进行语义分割

显着性检测 (SD) 方法通过具有像素级注释[14] 或图像级注释[15] 的外部显着性数据集生成区分图像前景和背景的显着性图。早期的显着性检测方法使用低级特征和启发式先验[16-17] 来检测对复杂场景不稳健的显着性对象。最近，基于深度学习的方法取得了显着的性能改进。比如，有的工作提出了一个深层层次网络来学习粗略的全局显着图，然后逐步细化它[18]。随着显著性图检测的性能提升，越来越多的工作开始利用显著性图进行语义分割，如一些工作利用仅仅使用显著图对图像进行弱监督语义分割[19]; 一些工作将显着图与特定类别的注意线索相结合，以生成可靠的伪掩码[20-21]; 有的工作用实例级别的显著性图去学习目标的相似性图形[22]; 有的工作使用一个单独网络共同解决了弱监督语义分割（WSSS）和显著性检测（SD）任务[23]。

3 本文方法

3.1 本文方法概述

文章提出了一种用于弱监督语义分割 (WSSS) 的新框架，称为显式伪像素监督 (EPS)。论文的关键点在于关键点在于结合两个互补的信息，即定位图和显著性图。本文使用显著性图作为定位图中目标和背景的伪像素反馈，设计了一个分类器，共有 $C+1$ 类，其中多了一类背景类。通过这个分类器，我们得到 $C+1$ 个定位图, 如图 1所示。

为了解决边界不匹配问题，论文从 C 个定位图中估计出一个前景图，然后和显著性的前景做匹配。这样带有目标标签的定位图就接收到了显著性图的伪像素反馈，改善了目标边界信息。为了解决来自非目标对象的共现像素问题，论文也将定位图的背景图和显著性图做匹配。EPS 的损失函数共分为两个部分: 显着图的显着性损失 $\mathcal{L}_{\text{sal}}$ 以及多标签分类损失 $\mathcal{L}_{\text{cls}}$

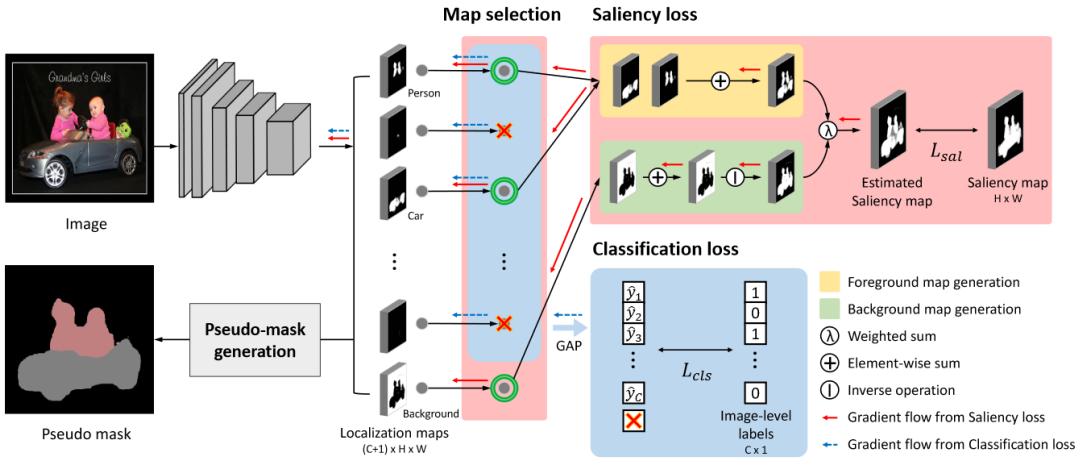

图 1: 总体框架图

3.2 显式伪像素监督

显着图的主要优点是提供对象轮廓，可以更好地揭示对象边界。为了利用这个属性，论文将显着图与两种情况相匹配：前景和背景。为了使类定位图与显着性图具有可比性，作者合并目标标签的定位图并生成前景图，$\mathbf{M}_{fg} \in \mathbb{R}^{H \times W}$。此外，还通过对背景图进行反转来表示前景，背景图是背景标签

$\mathbf{M}_{bg} \in \mathbb{R}^{H \times W}$ 的定位图。

具体来说，可以使用 $\mathbf{M}_{fg}$ 和 $\mathbf{M}_{fg}$；来估计显著性图 $\hat{\mathbf{M}}_s$:

$$\hat{\mathbf{M}}_s = \lambda \mathbf{M}_{fg} + (1-\lambda)(1-\mathbf{M}_{bg}) \tag{1}$$

其中 $\lambda \in [0,1]$ 是一个超参数，用于调整前景图和背景图的反转的加权和。

以前的工作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出的显着图不兼容。因为显着性模型学习不同数据集的统计数据，所以这种系统误差是不可避免的。除非考虑这个错误，否则相同的错误可能会传播到模型并导致性能下降。

为了解决系统误差，作者利用定位图和显着图之间的重叠率制定了一种有效的策略。具体来说，如果 $\mathbf{M}_i$ 与显着图重叠超过 τ%，则第 i 个定位图 $\mathbf{M}_i$ 被分配给前景，否则为背景。前景和背景图通过以下方式计算：

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

其中，$y \in \mathbb{R}^C$ 是二值图像级标签，$\mathcal{O}\left(\mathbf{M}_i, \mathbf{M}_s\right)$ 是计算 $\mathbf{M}_i$ 和 $\mathbf{M}_s$ 之间重叠率的函数。为此，我们首先对定位图和显着性图进行二值化，使得：对于像素 p，如果 $\mathbf{M}_k(p) > 0.5$，则 $\mathbf{B}_k(p) = 1$，否则 $\mathbf{B}_k(p) = 0$。$\mathbf{B}_i$ 和 $\mathbf{B}_s$ 分别为 $\mathbf{M}_i$ 和 $\mathbf{M}_s$ 对应的二值化图。然后计算 $\mathbf{M}_i$ 和 $\mathbf{M}_s$ 之间的重叠率，即 $\mathcal{O}\left(\mathbf{M}_i, \mathbf{M}_s\right) = |\mathbf{B}_i \cap \mathbf{B}_s| \,/\, |\mathbf{B}_i|$。此外，作者根据实验设置 $\tau = 0.4$。

作者将背景标签的定位图与未被选为前景的定位图结合起来，而不是背景标签的单一定位图。虽然简单，但可以绕过显着图边界模糊的缺陷，有效地训练一些被显着图忽略的对象。

3.3 联合训练

EPS 的整体训练目标由两部分组成，即显着图的显着性损失 $\mathcal{L}_{\text{sal}}$ 以及多标签分类损失 $\mathcal{L}_{\text{cls}}$。首先，显着性损失 $\mathcal{L}_{\text{sal}}$ 是通过测量实际显着性图 $\mathbf{M}_s$ 和估计的显着性图 $\hat{\mathbf{M}}_s$ 之间的平均像素级距离来制定的。

$$\mathcal{L}_{sal} = \frac{1}{H \cdot W} \left\| \mathbf{M}_s - \hat{\mathbf{M}}_s \right\|^2 \tag{3}$$

其中 $\mathbf{M}_s$ 是从现成的显着性检测模型，在 DUST 数据集[24] 上训练的PFAN[25] 获得的。

分类损失是通过图像级标签 y 及其预测值 $\hat{y} \in \mathbb{R}^C$ 之间的多标签软边缘损失来计算的，这是每个目标类在定位图上的全局平均池化的结果。

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

其中 $\sigma(\cdot)$ 是 sigmoid 函数。总的训练损失是多标签分类损失和显着性损失的总和，即 $\mathcal{L}_{\text{total}} = \mathcal{L}_{cls} + \mathcal{L}_{\text{sal}}$

4 复现细节

4.1 与已有开源代码对比

本次实验的 EPS 框架代码是开源的。但为了进一步优化生成的掩码的质量，参考了 AffinityNet 的工作[26], 我们针对该工作重新训练了一个语义亲和网络，如图 2所示。

与 AffinityNet[26]不同是，我们是对 EPS 框架产生最终掩码进行语义传播游走，而原工作是对特征提取后的 CAM 图进行亲和度计算。具体来说，将训练样本生成的掩码来生成语义标签，并放入到 AffinityNet 中让网络学习其亲和度。考虑到计算效率，AffinityNet 用于预测卷积特征映射时，其中一对特征向量之间的语义相似度是由他们之间的 L1 距离决定的。特征 i 和 j 之间的语义亲和度由 $\mathbf{W_{ij}}$ 表示并定义为:

$$W_{ij} = \exp\left\{-\left\|f^{\text{aff}}\left(x_i, y_i\right) - f^{\text{aff}}\left(x_j, y_j\right)\right\|_1\right\} \tag{5}$$

AffinityNet 以梯度下降的方式训练。考虑到缺乏上下文信息和计算量两个因素，样本提取的过程中只考虑充分相邻的像素对。因此由一个像素对组成的训练样本 $\mathcal{P}$ 表示如下：

$$\mathcal{P} = \{(i, j) \mid \mathrm{d}\left(\left(x_i, y_i\right), \left(x_j, y_j\right)\right) < \gamma, \forall i \neq j\} \tag{6}$$

其中 $\mathrm{d}(\cdot,\cdot)$ 是欧氏距离，γ 是限制所选对之间距离的搜索半径。接着利用 AffinityNet 生成对应的概率矩阵在掩码上进行随机游走，将定位图的激活分数传播到同一语义实体的附近区域，来增强其边界清晰度。

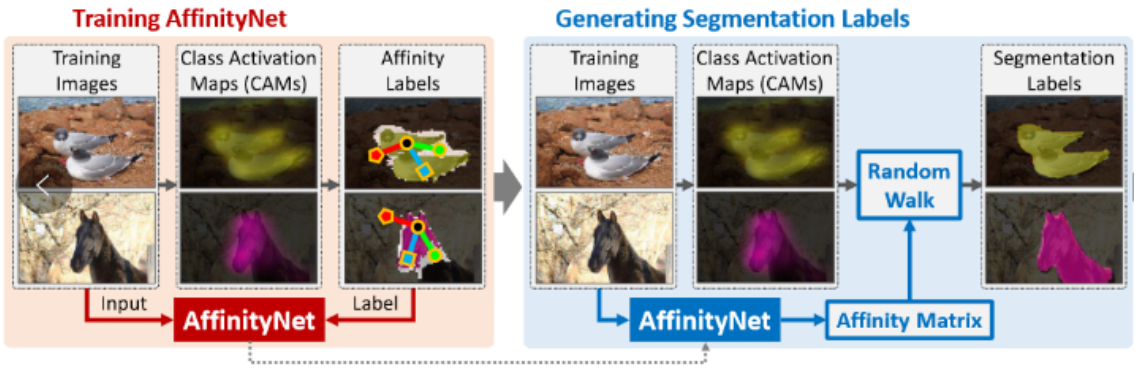

图 2: 定位图优化流程

4.2 实验环境搭建

本次实验运行在 Linux 上，所需要的相关软件和库有：Python3.6、CUDA10.1、pytorch1.8、torchvision0.9.1、MXNet、Pillow 和 opencv-python 等。

数据集来源于 PASCAL VOC 2012[27], 由 21 个类（即 20 个对象和背景）组成. 分别有 1,464、1,449 和 1,456 张图像用于训练、验证和测试集。按照语义分割的惯例，我们使用包含 10,582 张图像的增强训练集[28]。我们使用 PASCAL VOC 2012 上的验证集和测试集来评估我们的方法, 采用平均交并比 (mIoU) 来衡量分割模型的准确性。

掩码的生成和优化、掩码作为标签进行分割，均在 4 张 NVIDIA GeForce RTX 3090 上完成。

4.3 实施细节

对于 EPS 网络，我们选择 ResNet38[29]作为我们方法的骨干网络，输出步长为 8。所有骨干模型都在 ImageNet[30]上进行了预训练。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成 448 × 448。对于分割网络，我们采用 DeepLab-LargeFOV(V1)[31]和 DeepLab-ASPP(V2)[32]。

对于 Affnitynet，主干网络选择 DNNs，是 ResNet38[29]的修改版本。为了得到主干网络，首先去除了原始模型的最终 GAP 和全连接层。然后将最后三层的卷积层替换为输入步幅为 1 的空洞卷积，并调整它们的扩张率，以便主干网络返回步幅为 8 的特征图。

Affnitynet 旨在聚合骨干网络的多级特征图，以便在计算亲和力时利用在各种视野中获取的语义信息。为此，选择了骨干网络最后三层输出的特征图。在聚合之前，对于第一、第二和第三个特征图，它们的通道维度分别通过单独的 1×1 卷积层减少到 128、256 和 512。然后将特征图连接成具有 896 个通道的单个特征图。我们最后在顶部添加了一个具有 896 个通道的 1×1 卷积层以进行自适应

5 实验结果分析

图 3展示了实验结果的部分可视化，可以看到，EPS 的分割结果对于比 groundtruth 十分相近。并且在多物体以及物体边缘处都取得比较好的分割结果。

图 3: PASCAL VOC 2012 分割结果图。(a) 输入图像，(b) groundtruth 和 (c) EPS。

将 EPS 与初始 baseline CAM[33]和三种最先进的方法，即 SEAM[34]、ICD[35] 和 SGAN[36], 结果如图 4所示。此外，基于 ResNet101 的 DeepLabV1 和 DeepLabV2 的模型，在 PASCAL VOC 2012 数据集中实现了最优的性能，如表 1所示。

Method	Seg	sup	val
ICD[35]$_{CVPR'20}$	V1	I	64.1
SC-CAM[37]$_{CVPR'20}$	V1	I	66.1
MCIS[4]$_{ACCESS'20}$	V2	I.+S.	66.2
SGAN[36]$_{ACCESS'20}$	V2	I.+S.	67.1
ICD[35]$_{CVPR'20}$	V1	I.+S.	67.8
Our EPS	V1	I.+S.	**71.0**
Our EPS	V2	I.+S.	70.9

表 1: PASCAL VOC 2012 上的分割结果 (mIoU)。所有结果均基于 ResNet101

此外，我们复现工作与使用 Affnity Net 优化后的结果如表 2所示。对比自己的复现效果和加入 Affnity Net 优化后的性能，在 DeeplabV1 下取得了 1.23% 性能提升，在 DeeplabV2 下提升了 1.67%。对比论文中的 71.0% 同为 DeeplabV2 情况下，性能高了 0.48%，比 DeeplabV1 高了 0.38%。但是，可以看到我们对于论文方法的复现效果仍然还有差距：在 DeeplabV1 下相差 1.78%，在 DeeplabV2 下相差 2.19%。

Method	Paper	ours	w/ Affinity Net
EPS（v1+101）	71.0	69.22	70.46
EPS（v2+101）	70.9	69.71	**71.38**

表 2: 论文结果、复现结果以及优化结果比对

总的来说，加入 Affinity Net 语义亲和网络的框架性能相对于基础框架都有大幅度的提升。这样

证明了 Affinity Net 的正确性和优越性：通过估计图中连通像素对的语义相似度。针对每一个类别，在掩码中通过随机游走策略传播到周围语义相同的区域。

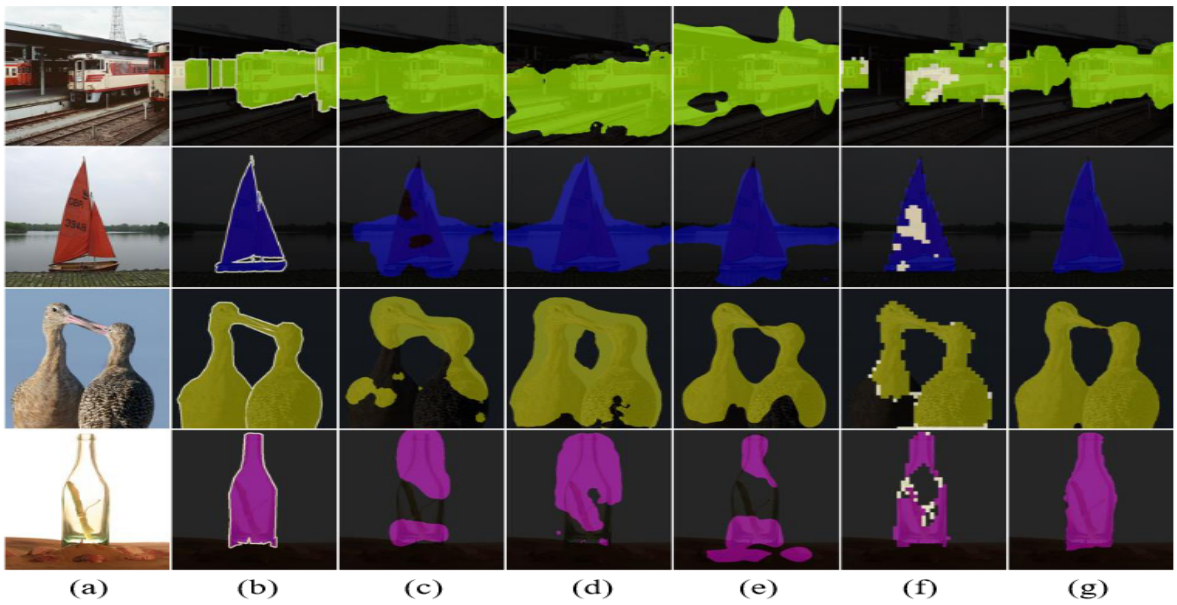

图 4: PASCAL VOC 2012 上伪掩码的定性比较。(a) 输入图像，(b) groundtruth，(c) CAM，(d) SEAM，(e) ICD，(f) SGAN 和 (g) EPS。

6 总结与展望

这次实验主要的工作是复现 EPS 网络框架，以及针对性的加入了 Affinity Net 进行进一步的优化。对比于自己的复现结果，加入 Affnity Net 后 EPS 网络性能在不同的分割网络下都有了一定的提升。但基础框架的复现对比于论文中的结果还存在一定差距。出现这个情况的主要原因是 Deeplab 的分割性能与它的超参数有很强的关联，论文作者也没有给出相关的分割模型超参数让我们完美的复现。通过一个多月的数百次参数调整，以及其他性能提升的 trick 加入，最终也比论文中的效果低了一点。

未来将会进一步探索性能不如原文的原因，使得性能尽可能的与原文中的效果相同。此外，除了加入 Affnity Net 进行优化之外，还打算从 EPS 本身的网络框架入手，针对边缘共像素难以解决的问题针对性的设计损失函数。并且考虑加入原型和对比学习的方法，将每个像素分配到一个置信度较高的原型中，与同为一个原型的像素拉近距离，与不同为一个原型的像素拉远距离。而且原型之间也可以相互疏远，以便更好的区分每一个物体。

通过这次实验，让我从研一的科研小白逐渐变成能够发现问题，提出解决办法、搭建网络、训练网络、调整参数的科研入门人员。科研并不是一帆风顺的，在这次实验中我也遇到了很多挫折。比如一开始的环境搭建失败，到后来的原文复现效果较差，以及优化后的性能不佳等问题。但凭借着对科研的热情，这些问题都一一被解决。每当解决一个问题后，心里的自豪和满足感都会鼓励我继续往前走，这或许就是科研的魅力所在。这门课程真的大大提升了我的思考能力，编程能力以及心理素质，对于刚入门的研一同学来说是十分可贵和难得的。希望这门课程能够继续延续下去，也希望自己的在科研的高峰上越攀越高。

参考文献

[1] LEE S, LEE M, LEE J, et al. Railroad is not a Train: Saliency as Pseudo-pixel Supervision for Weakly Supervised Semantic Segmentation[Z]. 2021.

[2] FAN J, ZHANG Z, TAN T, et al. CIAN: Cross-Image Affinity Net for Weakly Supervised Semantic Segmentation[C]//National Conference on Artificial Intelligence. 2020.

[3] LIU Y, WU Y H, WEN P, et al. Leveraging Instance-, Image- and Dataset-Level Information for Weakly Supervised Instance Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(3): 1415-1428. DOI: 10.1109/TPAMI.2020.3023152.

[4] SUN G, WANG W, DAI J, et al. Mining Cross-Image Semantics for Weakly Supervised Semantic Segmentation[C]//ECCV. 2020.

[5] QI X, LIU Z, SHI J, et al. Augmented Feedback in Semantic Segmentation Under Image Level Supervision[C]//LEIBE B, MATAS J, SEBE N, et al. Computer Vision – ECCV 2016. Cham: Springer International Publishing, 2016: 90-105.

[6] DONG Z, HANWANG Z, JINHUI T, et al. Causal Intervention for Weakly Supervised Semantic Segmentation[C]//NeurIPS. 2020.

[7] KUMAR SINGH K, JAE LEE Y. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 3524-3533.

[8] LI K, WU Z, PENG K C, et al. Tell me where to look: Guided attention inference network[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 9215-9223.

[9] WEI Y, FENG J, LIANG X, et al. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1568-1576.

[10] AHN J, KWAK S. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4981-4990.

[11] HUANG Z, WANG X, WANG J, et al. Weakly-supervised semantic segmentation network with deep seeded region growing[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7014-7023.

[12] WEI Y, XIAO H, SHI H, et al. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7268-7277.

[13] LEE S, LEE J, LEE J, et al. Robust tumor localization with pyramid grad-cam[J]. arXiv preprint arXiv:1805.11393, 2018.

[14] HOU Q, CHENG M M, HU X, et al. Deeply supervised salient object detection with short connections [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 3203-3212.

[15] WANG L, LU H, WANG Y, et al. Learning to detect salient objects with image-level supervision[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 136-145.

[16] YANG C, ZHANG L, LU H, et al. Saliency detection via graph-based manifold ranking[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 3166-3173.

[17] JIANG Z, DAVIS L S. Submodular salient region detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 2043-2050.

[18] LIU N, HAN J. Dhsnet: Deep hierarchical saliency network for salient object detection[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 678-686.

[19] WEI Y, LIANG X, CHEN Y, et al. Stc: A simple to complex framework for weakly-supervised semantic segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(11): 2314-2320.

[20] WANG X, YOU S, LI X, et al. Weakly-supervised semantic segmentation by iteratively mining common object features[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1354-1362.

[21] YAO Q, GONG X. Saliency guided self-attention network for weakly and semi-supervised semantic segmentation[J]. IEEE Access, 2020, 8: 14413-14423.

[22] FAN R, HOU Q, CHENG M M, et al. Associating inter-image salient instances for weakly supervised semantic segmentation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 367-383.

[23] ZENG Y, ZHUGE Y, LU H, et al. Joint learning of saliency detection and weakly supervised semantic segmentation[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 7223-7233.

[24] WANG L, LU H, WANG Y, et al. Learning to detect salient objects with image-level supervision[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 136-145.

[25] ZHAO T, WU X. Pyramid feature attention network for saliency detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 3085-3094.

[26] AHN J, KWAK S. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4981-4990.

[27] EVERINGHAM M, ESLAMI S, VAN GOOL L, et al. The pascal visual object classes challenge: A retrospective[J]. International journal of computer vision, 2015, 111(1): 98-136.

[28] HARIHARAN B, ARBELÁEZ P, BOURDEV L, et al. Semantic contours from inverse detectors[C]//2011 international conference on computer vision. 2011: 991-998.

[29] WU Z, SHEN C, VAN DEN HENGEL A. Wider or deeper: Revisiting the resnet model for visual recognition[J]. Pattern Recognition, 2019, 90: 119-133.

[30] DENG J, DONG W, SOCHER R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern recognition. 2009: 248-255.

[31] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Semantic image segmentation with deep convolutional nets and fully connected crfs[J]. arXiv preprint arXiv:1412.7062, 2014.

[32] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(4): 834-848.

[33] ZHOU B, KHOSLA A, LAPEDRIZA A, et al. Learning deep features for discriminative localization[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2921-2929.

[34] WANG Y, ZHANG J, KAN M, et al. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 12275-12284.

[35] FAN J, ZHANG Z, SONG C, et al. Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 4283-4292.

[36] YAO Q, GONG X. Saliency guided self-attention network for weakly and semi-supervised semantic segmentation[J]. IEEE Access, 2020, 8: 14413-14423.

[37] CHANG Y T, WANG Q, HUNG W C, et al. Weakly-supervised semantic segmentation via sub-category exploration[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 8991-9000.