Integral Object Mining via Online Attention Accumulation*

Peng-Tao Jiang[1] Qibin Hou[1] Yang Cao[1] Ming-Ming Cheng[1†]
Yunchao Wei[2] Hong-Kai Xiong[3]
[1]TKLNDST, CS, Nankai University [2]UTS [3]Shanghai Jiaotong University
pt.jiang@mail.nankai.edu.cn cmm@nankai.edu.cn

摘要

通过图像分类器生成的目标注意力图通常用作弱监督分割方法的先验。然而，一般的图像分类器产生的注意力只在最有区分度的目标部分，这限制了弱监督分割任务的性能。因此，如何以弱监督的方式有效地识别整个目标区域一直是一个具有挑战性并且有意义的问题。我们观察到，分类网络产生的注意力图在训练中持续关注不同的目标部分。为了积累已发现的不同目标部分，我们提出了一个在线注意力积累策略（*OAA*），该策略为每个训练图像中的每个目标类别分别维护了一个累积注意力图，随着训练进行，整体目标区域的效果可以逐步提升。这些累积注意力图又充当了像素级别的监控，能够进一步辅助网络发现更多的完整目标区域。我们的方法（*OAA*）可以插入到任何分类网络中，并且随着训练的进行，逐步将可区分的区域积累为完整目标。尽管它很简单。当将得到的注意力图运用于弱监督的语音分割任务时，我们的方法改进了在*PASCAL VOC 2012*分割基准测试上现有的最新方法，在测试集上达到了*66.4%*的*mIoU*分数。代码位于*http://mmcheng.net/oaa/*.

1. 引言

受益于大规模的像素级训练数据和先进的卷积神经网络结构，全监督语义分割方法，诸如，当前已取得巨大进展。然而，建立一个大规模的像素级精确的数据集是十分昂贵的，并且需要相当的人力和时间。为了节约人工劳动，研究者打算使用弱监督的方式学习语义分割。例如[4, 21, 23, 39, 43]，边界框[28]，多点[2]甚至图像级别的标注[27]。在这些弱监督方式中，图像级别的标注比其余方式更加容易。因此，这篇论文中我们重点关注在图像级别监督下的语义分割。

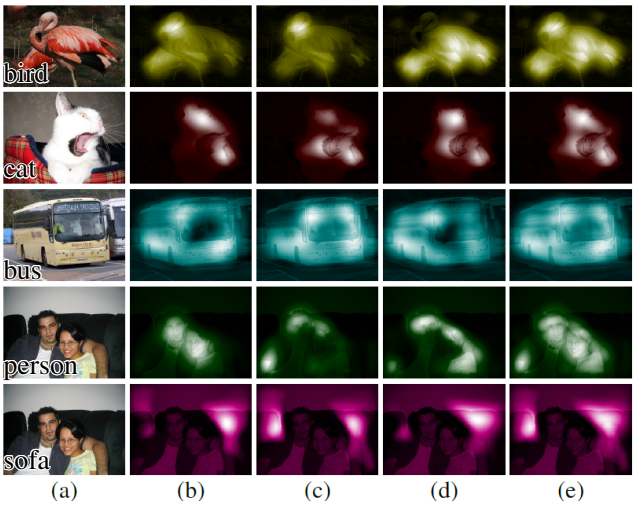

图 1. 对我们提出的方法的观察。(a)原图像(b-d)由分类网络在不同训练阶段生成的中间注意力图(e)通过简单地元素最大运算组合(b),(c)和(d)的注意力图得到的累积注意力图。可以轻松的观察到，有区分的区域在语义区域的不同部分连续移动。相比于(b)(c)(d)，在(e)中融合后的注意力图可以记录大部分语义区域。获得了在颜色上的最好观感。

由于分类模型在发现有区分度的注意力区域时

*本文为ICCV 2019论文[16]的中文翻译版。
†M.M. Cheng为通讯作者。

的良好能力，分类模型[30, 44]已被广泛应用于弱监督语义分割任务中，以生成特定类别的初始种子。但是，这些被发现的区域经常集中于语义对象中的一小部分，这限制了分割网络学习复杂像素级语义知识的能力。后来的方法考虑利用对抗擦除策略[36, 41]以挖掘更多的语义区域。不幸的是，随着训练过程的进行，有区分度的区域拓展，随之一些期望之外的背景同样被推测为前景。在[38]中，为了生成注意力，膨胀后的卷积被再次访问。然而，更大膨胀率的卷积层常常导致噪声区域的出现。

上述方法的一个共同点是他们都使用了最终分类模型以生成注意力图。在这篇文章中，我们从一个新的角度考虑注意力的生成。我们观察到，在分类网络收敛之前，不同训练阶段探查到的有区分度的区域会在语义对象的不同部分上不断漂移。主要原因概括如下：

- 首先，强大的分类网络通常会针对特定类别寻求鲁棒的通用模式，以便可以很好的识别出该类别中的所有图像。因此，那些难以正确分类的训练样本使得在选择通用模式时发生更改，导致关注区域连续移动直到网络收敛。
- 其次，在训练期间，当前注意力模型生成的注意力图主要被之前的输入图像影响。因此，不同内容的图像和训练图像的不同顺序均会导致在中间注意力图中可区分的区域的变化。

更有趣的是，我们还观察到在不同训练阶段发现的区分区域通常是互补的，这反映了利用中间注意力图来检测整体对象的重要性。图 1（b-d）清楚的说明了这种现象，他显示了训练过程中注意力区域的变化。如果这些有区分度的区域能在中间注意力图中被记录，我们也许仅通过图像级别的监督就可以成功地提高检测完整语义对象的能力。

基于上述观察，我们介绍一种简单却有效的注意力生成方法，他可以考虑分类网络的中间状态。具体来说，我们提供了一个在线注意力累积（OAA）策略，其中维护了每个图像中每个类别的累积注意力图，以在不同训练阶段顺序积累分类网络产生的有区分度区域。中间注意力图的互补性使发现整体语义对象成为可能（见图 1（e））。尽管与CAM[44]相比，OAA生成的注意力图相对完整，一些目标区域的注意力值仍不够强。为了改善这种情况，我们进一步设计了一种混合损失函数（结合了增强损失和限制损失），以累积注意力图为软标签来训练累积注意力模型。这样，一个新的注意力模型将提升OAA策略并可以更完整的目标区域。为了评估我们方法生成的注意力图的质量，我们进行了一系列消融实验并将其应用于弱监督语义分割任务。我们在流行的PASCAL VOC 2012数据集分割基准[8]上展现了对现有方法的显著改进（测试集上的平均IoU分数为66.4%）。我们希望OAA的思想能够促进注意力模型，甚至其他领域的发展。

2. 相关工作

在这一部分，我们简要地回顾注意力模型的历史并描述与我们工作密切相关的弱监督语义分割方法。

2.1. 视觉注意力

迄今，许多以获得高质量的注意力为目标的方法被提出。作为早期的尝试，Simonyan等人[32]使用了错误反向传播策略以可视化语义区域。后来，CAM[44]通过将其用于卷积神经网络以检测类激活图来显示全局平均池化（GAP）层的功能。Grad-CAM[30]基于CAM，提出了一种通过将梯度流入最终卷积层以生成粗略注意力图，来为任何目标概念产生视觉解释的技术（例如图像分类，VQA与图像描述）。此外，一些研究人员收到自上而下的人类视觉注意力系统的启发，提出了一种名为激发反向传播[40]（Excitation Backprop）的方法。该方法通过概率的赢者通吃模型在网络中层级向下地传播自顶向下的信号。近来，不同于上述解释网络的方法，一些工作[14, 20, 38, 41, 42]通过定位语义对象大而完整的相关区域来进行弱监督语义分割并产生注意力图。所有上述方法均使用了最终分类模型以生成注意力。除了自上而下的视觉注意力外，最近的研究[13, 35, 38]也发现自底向上的显著性物体提示[6, 12, 34]对提取背景提示非常有用。

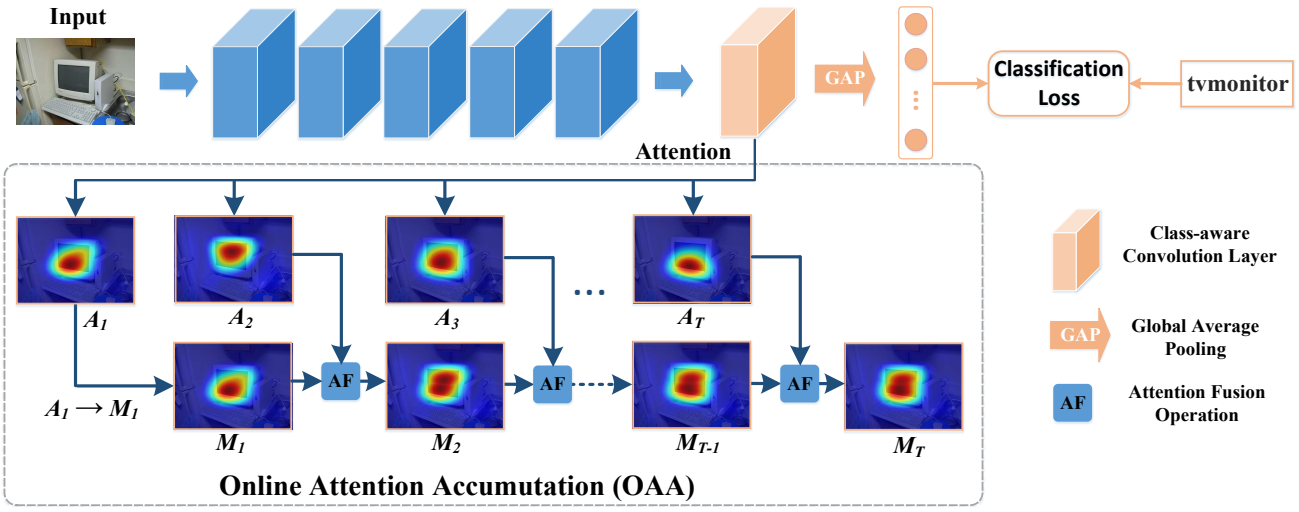

图 2. 我们在线注意力累积（OAA）过程的示意图。注意力图是从类感知卷积层生成的。OAA，在不同训练阶段，利用这些有区分度的注意力图的区域并通过一个简单的注意力融合策略将它们逐步集成到累积注意力图中。

2.2. 弱监督语义分割

由于诸多方法的提出，弱监督语义分割也取得了长足的进步。在这些方法中，我们仅介绍与我们工作密切相关的带图像级别监督的分割方法。主流方法[1, 15, 19, 35, 38]使用注意力图作为初始种子。典型地，SEC介绍了3种损失函数，分别名为种子损失、扩展损失和边界限制损失以拓展初始种子并同时训练分割模型。然而，由于目标相关的种子只覆盖较小且稀疏的语义区域，这些方法的性能均受到限制。

最近，研究人员提出了诸多基于分类网络的整体目标区域挖掘算法。在[36]，Wei等人提出了一种使用对抗擦除策略（AE-PSL）来逐步挖掘目标的不同区域以获得密集注意力图。但是，AE-PSL的过程相当复杂，需要重复训练并学习多个分类模型以获得不同的对象区域。GAIN[20]通过使用注意力图来提供自我指导，从而迫使网络将注意力集中于目标整体上，从而改进了对抗擦除策略。

3. 方法

在这个部分，我们描述了我们提出方法的流程并详细地解释了框架中每个组件的工作机制。图3阐释了我们方法的整个框架。

3.1. 注意力生成

在这篇文章中，我们采用了CAM[44]作为我们的默认有区分度的区域生成器。为了在训练阶段获得注意力图，我们通过最后一个卷积层输出的特定于类的特征图来生成注意力图，这种方式被证明[41]与CAM中的注意力生成过程相同。

基本构架可以在 2顶部找到。与之前工作[38, 41]类似，我们也采用了VGG16[33]作为骨干。首先，3个卷积层被加入到全卷积骨干的顶部，每个卷积层之后是用于非线性转换的ReLU层。一个内核大小为1×1的C个通道的类感知卷积层被添加到内核以进行注意力捕获。这里C是类别数。令F为类感知卷积层的输出。考虑到一些图片可能具有多个类别，我们将整个训练过程视为C的二分类问题。可以通过以下公式计算预测目标类别c的概率

$$p^c = \sigma(GAP(F^c)) \tag{1}$$

这里GAP是全局平均池运算，$o(\cdot)$是S形函数。交叉熵损失用于优化整个网络。为了获得给定图像I的注意力图，首先将特征图F输入ReLU层，然后执行一个简单的归一化操作以确保注意力图中的数值在0到1之间。

$$A^c = \frac{\mathrm{ReLU}(F^c)}{\max(F^c)} \tag{2}$$

我们将不同训练阶段生成地注意力图应用在OAA过

程。

3.2. 在线注意力累积

为了有效地进行我们的观察，我们提出了一个在线注意力累积（OAA）策略。当在不同的训练时期，将训练图像输入网络时，OAA结合从分类模型中生成的注意力图。具体来说，正如图 2所示，对于给定训练图像I中的每个目标类别c，我们建立一个累积注意力图M^c，用于保存已发现的有区分度的区域。我们的OAA首先在第一个时期使用类c的注意力图A_1[1]（即当训练图像首次输入到网络时获得A_1）来初初始化累积注意力图M_1。然后，当图片第二次输入网络时，OAA根据如下融合策略，通过结合M_1和最新生成的注意力图A_2更新累积注意力图。

$$M_2 = \mathrm{AF}(M_1, A_2) \tag{3}$$

这里$AF(\cdot)$代表注意力融合策略。相似地，在第t个时期，OAA使用注意力图A_t更新累积注意力图M_{t-1}，得到

$$M_t = \mathrm{AF}(M_{t-1}, A_t) \tag{4}$$

OAA不断重复上述更新过程，直到分类模型收敛，我们可以得到最终累积注意力图。在上述更新过程中，注意力融合策略负责保持这些中间注意力图中的有区分度的区域以构建更加完整的目标区域。

关于融合策略，我们提出了一个有效且简单的策略，即逐元素最大操作。它采用注意力图A_t和当前累积注意力图M_{t-1}之间的最大注意力值，其公式如下：

$$M_t = \mathrm{AF}(M_{t-1}, A_t) = \max(M_{t-1}, A_t) \tag{5}$$

采用最大化融合策略的OAA可以有效地将不同的有区分度的区域保存到累计注意力图中。如图5所示，相比于CAM生成的注意力图，OAA生成的累积注意力图有更多更完整的区域。我们也探索了OAA的平均融合策略。但是，与最大融合策略相比，mIoU的分数下降了1.6%。在4.3节，我们进行消融实验以显示两种融合策略间的差异。

[1]这里我们为了方便，忽略了类c。

值得一提的是，由于分类模型较弱并且可能在训练过程开始时关注噪声区域，我们使用目标类别的预测概率来决定是否积累响应注意力图。特别地，如果目标类别的分类得分高于所有非类别的得分，则我们会在OAA中累积目标类别的注意力图。否则，我们将放弃此注意力图以避免噪声。

3.3. 整体注意力学习

OAA整合了训练过程中不同时期的注意力图以生成更完整的目标区域。然而，OAA的弱点在于，分类模型自身无法增强一些低注意力值的目标区域。考虑到这种情况，我们通过将累积注意力图作为监督来引入新的损失函数，以训练完整的注意力模型来进一步提升OAA的性能，我们称其为OAA$^+$。

具体来说，我们使用累积注意力图作为软标签，正如[37]中所作的。每个注意力值被视为此位置属于对应目标类别的概率。我们采用图 2中所示的分类网络作为整体注意力模型，其中取消了全局平均池化层和分类损失。给定类别感知卷积层生成的得分图$\hat{F}$，位置j为某个类别c的概率可以表示为$q_j^c = \sigma(\hat{F}_j^c)$，其中$\sigma$是sigmoid函数。因此，类$c$中使用的多标签交叉熵损失[37]可以写成：

$$-\frac{1}{|N|}\sum_{j\in N}\left(p_j^c \log\left(q_j^c\right) + \left(1-p_j^c\right)\log\left(1-q_j^c\right)\right) \tag{6}$$

这里p_j^c表示标准化后的累积注意力图的值。优化后，增强注意力图可以直接从类感知卷积层获得。然而，利用上述交叉熵损失函数，产生的注意力图区域部分覆盖语义对象区域。原因是等式 6中的损失函数倾向于将具有低的特定类别注意力值的像素分类为类c的背景。

考虑到以上讨论，我们提出了一种改进的混合损失。给定类c的累积注意力图（值域从0到1），我们首先将其划分为软增强区域N_+^c与软约束区域N_-^c，其中N_-^c包括$p_j^c = 0$的像素，N_+^c包含其他像素。对于像素集N_+^c，我们移除等式 6的最后一项以进一步提升注意力区域但不抑制第注意力值区域。形式上，

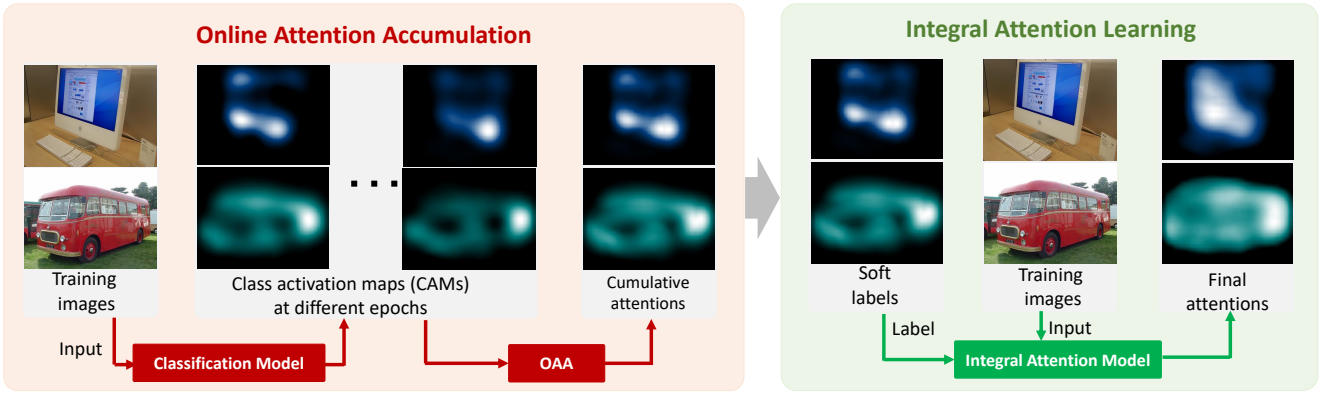

图 3. OAA$^+$方法的流程。将在不同训练阶段生成的注意力图融入累积注意力图以尽可能完整地挖掘目标区域。然后得到的累积注意力图用作像素级监督以训练整体注意力模型，这进一步提升了注意力图的质量。

我们有N_+^c的损失函数为

$$\mathcal{L}_+^c = -\frac{1}{|N_+^c|}\sum_{j\in N_+^c} p_j^c \log\left(q_j^c\right) \tag{7}$$

由于这里仅给出图像级标签且语义对象形状不规则，累积注意力图中的注意力区域常常包含非目标像素。因此，在等式 7中，我们使用p_j^c而不是1作为客观事实标签，这样在累积注意力图中非语义区域的低注意力值几乎对网络没有负面影响。对于$p_j^c = 0$的N_-^c，等式 6中的损失函数成为如下形式：

$$\mathcal{L}_-^c = -\frac{1}{|N_-^c|}\sum_{j\in N_-^c} p_j^c \log\left(1-q_j^c\right) \tag{8}$$

结果，我们的整体注意力模型的总损失函数可以通过如下公式计算：

$$\mathcal{L} = \sum_{c\in C}\left(\mathcal{L}_+^c + \mathcal{L}_-^c\right) \tag{9}$$

这样，根据等式 7、等式 8中的损失函数（这些损失函数限制了注意力区域向背景的过度扩张），软增强区域中较低的值也有助于优化。

基于提出的损失函数，我们可以训练一个整体注意力模型以进一步增强目标区域的较低注意力值。在推论时，可以从整体注意力模型的类感知卷积层直接获得改进的注意力图。另外，图5展示了我们注意力图的一些可视化结果，更定量的分析在4.3节中。

4. 实验

为了说明我们方法的有效性，我们将OAA与OAA$^+$生成的注意力图作为启发性线索应用于弱监督语义分割任务。我们使用注意力图与显著图[12]分别提取对象线索与背景线索。之后，这些线索用于生成伪分割标注。我们将与最大值对应的类别标签分配给代理分割标签中的像素。忽略所有冲突的像素以进行训练。由上述方法生成的代理真实数据被用于训练分割模型。在以下小节中，我们提供一系列消融研究并将我们的方法与之前最新方法进行了比较。

4.1. 数据集和设置

数据集与评判标准 我们在PASCAL VOC2012分割基准[8]上评估我们的方法，这个数据集包含20种语义类别与背景。正如大多数以前工作所作的，我们也使用增强训练集[9]训练。因此，我们共有10582张训练图像。在测试阶段，我们在验证与测试集上比较了我们的方法与之前的方法在平均交并比（mIoU）这个评判标准上的差异。由于测试集的分割标注并不公开，我们向PASCAL VOC评估服务器提交了我们的测试结果以获得评分。

网络设置 分类网络的超参数如下：mini-batch大小（5），权重衰减（2×10^{-4}），动量（0.9）。我们的初始学习率设置为1×10^{-3}，这个学习率在2×10^4次迭代后减少到原来的十分之一。我们总共运行分类网络进行3×10^4次迭代。我们使用除去全局平

均池化层与分类损失的分类网络作为整体注意力模型。整体注意力模型的超参数与分类网络中使用的一致。如大多数以前工作，我们使用DeepLab-LargeFOV模型[5]作为细分网络。分割网络以10个图像的小批量进行训练，并在1.5×10^4次迭代后停止。其他所有超参数均与[5]中相同。我们的报告基于VGG16[33]与ResNet-101[10]主干的结果。

4.2. 与先进方法的对比

在这个小节，我们将我们的方法与之前的仅依赖图像级标签的弱监督语义分割方法进行对比。表4列出了我们的方法与这些对比方法在测试集和验证集上的所有结果。不难观察到，无论使用哪个主干，我们方法的mIoU分数均比之前的最先进技术有所提升。在之前的最先进方法中，MIL[27]与WebS-i2[17]使用了更多的训练图像（分别是700K与19K）。另外，Hong等人[11]利用附加视频数据提供的丰富的时间动态信息，可帮助轻松地从视频中找出完整的语义对象。尽管仅使用了10K的图像数据，但我们的OAA方法在验证集上的得分比上述三个方法分别高了21.1%，9.7%和5.0%。这个事实很好地说明了，我们的整体注意力模型所产生的注意力图，对于目标对象的所有部分都可以有效地检测到的更完整的语义区域。

相比于AE-PSL[36]，我们的OAA方法，在无需训练多分类模型的情况下，获得了更高的mIoU得分（61.6% v.s. 55.0%）。并且，GAIN[20]以端到端的方式采用了一个自指导擦除策略。但我们的分割结果的mIoU得分比GAIN提升了7%以上（63.1% v.s. 55.3%）。与那些基于擦除方法的对比表明收集中间注意力图更为有效。在[38]中，Wei等人发现了膨胀卷积对发现完整目标的能力。然而，它通常会引入一些不相关的像素，因为大膨胀率的卷积经常集中在目标区域的外部。不同的是，我们的方法并不使用大膨胀率的卷积，因此可以弱化无关像素的影响。如表1中所示，我们的方法在验证集和测试集上都将[38]中的方法提升了2%。另外，我们也展示了基于ResNet-101[10]主干的分割结果。明显，我们提出的方法在PASCAL VOC 2012分割基准上得到了最好的结果。

方法	监督	验证集	测试集
主干: VGGNet [33]			
CCNN [26]	10K	35.3%	-
EM-Adapt [25]	10K	38.2%	39.6%
MIL [27]	700K	42.0%	-
DCSM [31]	10K	44.1%	45.1%
SEC [19]	10K	50.7%	51.7%
AugFeed [28]	10K	54.3%	55.5%
STC [37]	50K	49.8%	51.2%
Roy et al. [29]	10K	52.8%	53.7%
Oh et al. [24]	10K	55.7%	56.7%
AE-PSL [36]	10K	55.0%	55.7%
Hong et al. [11]	970K	58.1%	58.7%
WebS-i2 [17]	19K	53.4%	55.3%
DCSP [3]	10K	58.6%	59.2%
TPL [18]	10K	53.1%	53.8%
GAIN [20]	10K	55.3%	56.8%
DSRG [15]	10K	59.0%	60.4%
MCOF [35]	10K	56.2%	57.6%
Ahn et al [1]	10K	58.4%	60.5%
Wei et al [38]	10K	60.4%	60.8%
SeeNet [14]	10K	61.1%	60.7%
OAA (Ours)	10K	61.6%	61.9%
OAA$^+$ (Ours)	10K	63.1%	62.8%
主干: ResNet [10]			
DCSP [3]	10K	60.8%	61.9%
DSRG [15]	10K	61.4%	63.2%
MCOF [35]	10K	60.3%	61.2%
Ahn et al [1]	10K	61.7%	63.7%
SeeNet [14]	10K	63.1%	62.8%
OAA (Ours)	10K	63.9%	65.6%
OAA$^+$ (Ours)	10K	65.2%	66.4%

表 1. 与以前最先进算法在测试集与验证集上的定量对比。OAA$^+$表示注意力图是根据 3.3小节中描述的整体注意力模型生成的。

4.3. 消融分析

在这一节中，我们进行了一系列消融实验并

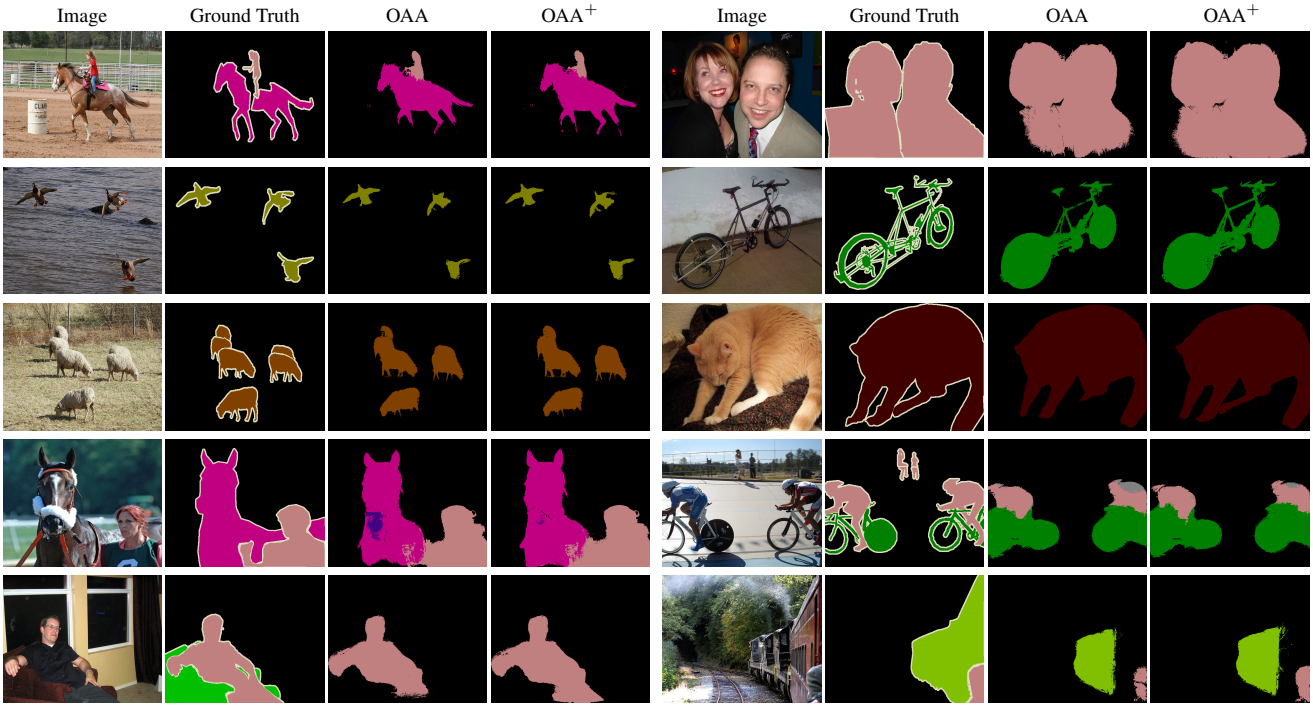

图 4. PASCAL VOC 2012验证集上分别使用OAA和OAA$^+$注意力图定性分割结果。我们在最后一行也展示了一些失败结果。

给出了我们提出方法有效性的具体分析。并且，我们说明了产生的注意力图是如何使语义分割任务受益。请注意，在此小节我们使用VGGNet版本的DeepLab-LargeFOV模型。

累积策略 注意力融合策略在OAA中被用于在不同时期的中间注意力图中累积已发现的有区分度的区域。除了最大融合策略，我们还研究了一种平均融合策略，可以表示为：

$$M_t = \frac{1}{t}((t-1)M_{t-1} + A_t) \tag{10}$$

如表 2中所示，在不使用OAA的情况下，使用CAM[44]的注意力在验证集上得出的mIoU分数为53.9%.当使用平均融合策略的OAA时，结果可以提高至57.0%.当使用最大融合策略代替平均融合策略时，我们的mIoU分数为58.6%，这大大提高了基于CAM[44]的结果。另外，我们观察到采用最大融合策略的OAA比采用平均融合策略的OAA更加有效。这是因为平均融合策略会平均中间注意力图中所有的注意力值，这降低了最终累积注意力图上的注意力值。因此，在下文中，我们将最大融合策略作为OAA的默认融合策略。请注意，这篇文章的目标是证明OAA的有效性，因此我们仅选择OAA逐元素最大融合策略。设计更为复杂的融合策略超出了这篇文章的范畴，但我们鼓励读者更进一步探索更为有效的融合策略。

OAA$^+$的损失函数 如3.3小节所述，累积注意力图后面被用作软标签以训练整体注意力模型来生成包含更完整更准确目标区域的注意力图。在表 2中，我们展示了使用不同损失函数的定量结果。可以看到，当将标准多标签交叉熵（MCE）[37]替换为我们提出的混合损失（HL）后，性能提升了8.4%。当使用多标签交叉熵时，输出注意力图常常覆盖较小的目标区域。相反，我们提出的混合损失可以进一步提升我们OAA生成的累积注意力图的质量。

不同策略的结果 除了可视化比较之外，我们还对PASCAL VOC 2012数据集进行了一系列消融实验。如表 2所示，我们展示了使用不同策略训练分割网络得到的注意力图的mIoU得分。在表 2的第三和最后一行，可以看到，使用OAA$^+$可以将OAA在验证集上结果进一步提升1.0%，这表明我们的整体注意力模型与提出的损失函数能够进一步帮助改善累积注意力图的质量。

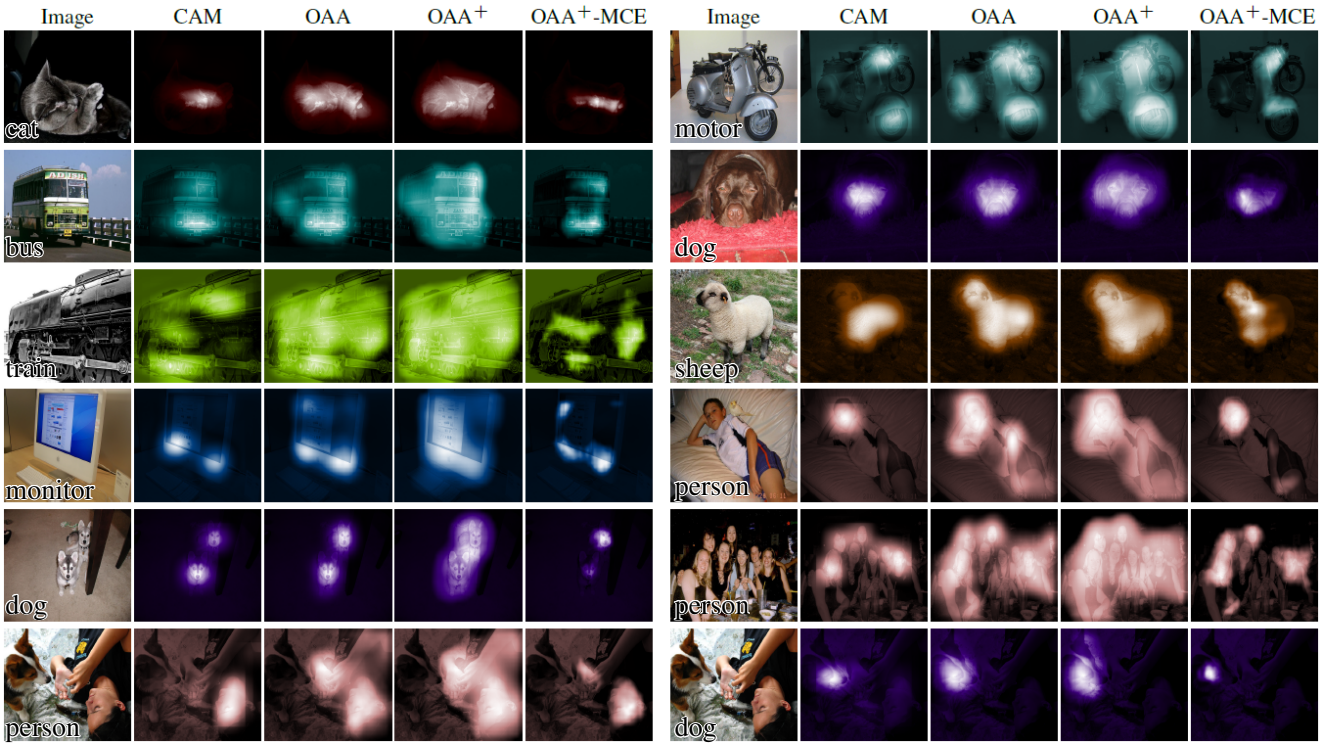

图 5. CAM[44]、OAA、OAA$^+$和OAA$^+$-MSC生成的注意力图之间的视觉比较。OAA$^+$和OAA$^+$-MSC分别表示通过等式 9中提出的混合损失和等式 6中的多标签交叉熵损失学习得到的整体注意力模型。

No.	AVE	MAX	MCE	HL	mIoU (val)
1					53.9%
2	✓				57.0%
3		✓			58.6%
4		✓	✓		51.2%
5		✓		✓	59.6%

表 2. 在PASCAL VOC 2012验证集上使用不同设置时mIoU分数对比。AVE：采用平均融合策略的OAA。MAX：采用最大融合策略的OAA。MSC：采用等式 6中多标签交叉熵的OAA$^+$。HL：采用等式 9中混合损失的OAA$^+$。

No.	#训练图像	比例	mIoU(val)
1	2,116	20%	54.6%
2	5,291	50%	57.3%
3	8,466	80%	58.9%
4	10,582	100%	59.6%

表 3. 在PASCAL VOC 2012验证集上使用不同数量的训练图像时的对比。请注意，图像是随机选择的。**比例**：训练图像的百分比。**训练图像**：训练图像的数量。

可视化对比 在本段中，我们展示了一些在PASCAL VOC 2012数据集[8]上的定性结果并分别给出了CAM[44]、OAA与OAA$^+$产生的相应注意力图作为可视化对比。如图5所示，图像包含了多种不同场景，诸如拥有不同规模的物体的、密集的物体的和多个类别物体的图像。从所有展示的例子中，当与CAM[44]生成的注意力图相比较时，我们的累积注意力图可以发现不同规模的近乎完整的目标对象。在第五行，展示了包含多个物体的图片。可以发现，在这种情况下我们的累积注意力图仍然可以覆盖大多数语义区域。在最后一行，我们展示了包含了多个类别的样例。明显，我们的累积注意力图可以成功区分不同类别并密集地检测目标。另外，相比于OAA生成的累积注意力图，OAA$^+$生成的注意力图可以发现更完整的目标区域。此外，我们还在图4中展示了一些分割结果。

训练图像的数量 为了进一步研究注意力图的质量，我们尝试使用不同数量的训练图像来训练

分割网络。我们使用OAA$^+$产生的注意力图来生成代理分割标注。如表 3所示，随着训练图像的增加，mIoU得分也逐渐提升。更有趣的是，当仅使用2116张图像时，我们的分割网络仍可以达到54.6%的表现，这个结果优于基于CAM[44]的分割结果。这间接地表明了我们的注意力图是高质量的并且有助于分割任务。

5. 总结

在这篇文章中，我们探索了一个名为OAA的简单有效的框架，以发现更多完整的目标区域。我们维护一系列的累积注意力图，以在训练阶段将分类网络生成的注意力图中不同的有区分度的区域保存下来。另外，我们利用累积注意力图作为软标签以训练完整注意力模型，来通过OAA增强注意力图。我们的方法易于遵循且可以简单地融入任何分类网络中，以完整地发现目标物体区域。全面的实验表明当弱监督分割任务应用我们的注意力图时，我们的分割网络比以往先进方法效果要更好。未来，我们打算在大规模数据集上进行实验，例如MS COCO[22]与ImageNet[7]。

致谢. 本研究受NSFC（61572264, 61620106008），国家青年拔尖人才支持计划与天津市自然科学基金（17JCJQJC43700, 18ZXZNGX00110）支持。Yunchao Wei得到IBM伊利诺伊州认知计算系统研究中心（C3SR）与ARC DECRA DE190101315的部分支持。

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 *CVPR*, 2018. 2072, 2075

[2] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei. What's the point: Semantic segmentation with point supervision. In *ECCV*, 2016. 2070

[3] A. Chaudhry, P. K. Dokania, and P. H. Torr. Discovering class-specific pixels for weakly-supervised semantic segmentation. *arXiv preprint arXiv:1707.05821*, 2017. 2075

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015. 2070

[5] 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 TPAMI*, 2017. 2075

[6] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu. Global contrast based salient region detection. *IEEE TPAMI*, 37(3):569–582, 2015. 2071

[7] J. Deng. A large-scale hierarchical image database. In *CVPR*, 2009. 2078

[8] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 2015. 2071, 2074, 2077

[9] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *ICCV*, 2011. 2074

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2075

[11] S. Hong, D. Yeo, S. Kwak, H. Lee, and B. Han. Weakly supervised semantic segmentation using web-crawled videos. In *CVPR*, 2017. 2075

[12] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr. Deeply supervised salient object detection with short connections. *IEEE TPAMI*, 41(4):815–828, 2019. 2071, 2074

[13] Q. Hou, P. K. Dokania, D. Massiceti, Y. Wei, M.-M. Cheng, and P. Torr. Bottom-up top-down cues for weakly-supervised semantic segmentation. *EMMCVPR*, 2017. 2071

[14] Q. Hou, P.-T. Jiang, Y. Wei, and M.-M. Cheng. Self-erasing network for integral object attention. In *NeurIPS*, 2018. 2071, 2075

[15] Z. Huang, X. Wang, J. Wang, W. Liu, and J. Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *CVPR*, 2018. 2072, 2075

[16] P.-T. Jiang, Q. Hou, Y. Cao, M.-M. Cheng, Y. Wei, and H.-K. Xiong. Integral object mining via online attention accumulation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2070–2079, 2019. 2070

[17] B. Jin, M. V. Ortiz Segovia, and S. Susstrunk. Webly supervised semantic segmentation. In *CVPR*, 2017. 2075

[18] D. Kim, D. Yoo, I. S. Kweon, et al. Two-phase learning for weakly supervised object localization. In *ICCV*, 2017. 2075

[19] A. Kolesnikov and C. H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *ECCV*. Springer, 2016. 2072, 2075

[20] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu. Tell me where to look: Guided attention inference network. In *CVPR*, 2018. 2071, 2072, 2075

[21] G. Lin, A. Milan, C. Shen, and I. Reid. Refinenet: Multi-path refinement networks with identity mappings for high-resolution semantic segmentation. In *CVPR*, 2017. 2070

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2078

[23] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 2070

[24] S. J. Oh, R. Benenson, A. Khoreva, Z. Akata, M. Fritz, and B. Schiele. Exploiting saliency for object segmentation from image level labels. In *CVPR*, 2017. 2075

[25] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille. Weakly-and semi-supervised learning of a dcnn for semantic image segmentation. *arXiv preprint arXiv:1502.02734*, 2015. 2075

[26] D. Pathak, P. Krahenbuhl, and T. Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *ICCV*, 2015. 2075

[27] P. O. Pinheiro and R. Collobert. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, 2015. 2070, 2075

[28] X. Qi, Z. Liu, J. Shi, H. Zhao, and J. Jia. Augmented feedback in semantic segmentation under image level supervision. In *ECCV*, 2016. 2070, 2075

[29] A. Roy and S. Todorovic. Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation. In *CVPR*, 2017. 2075

[30] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 2071

[31] W. Shimoda and K. Yanai. Distinct class-specific saliency maps for weakly supervised semantic segmentation. In *ECCV*, 2016. 2075

[32] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 2071

[33] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 2072, 2075

[34] J. Wang, H. Jiang, Z. Yuan, M.-M. Cheng, X. Hu, and N. Zheng. Salient object detection: A discriminative regional feature integration approach. *IJCV*, 123(2):251–268, 2017. 2071

[35] X. Wang, S. You, X. Li, and H. Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *CVPR*, 2018. 2071, 2072, 2075

[36] 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 *CVPR*, 2017. 2071, 2072, 2075

[37] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, Y. Zhao, and S. Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE TPAMI*, 39(11):2314–2320, 2017. 2073, 2075, 2076

[38] 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 *CVPR*, 2018. 2071, 2072, 2075

[39] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal. Context encoding for semantic segmentation. In *CVPR*, 2018. 2070

[40] J. Zhang, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff. Top-down neural attention by excitation backprop. In *ECCV*, 2016. 2071

[41] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. Huang. Adversarial complementary learning for weakly supervised object localization. In *CVPR*, 2018. 2071, 2072

[42] X. Zhang, Y. Wei, G. Kang, Y. Yang, and T. Huang. Self-produced guidance for weakly-supervised object localization. In *ECCV*, 2018. 2071

[43] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *CVPR*, 2017. 2070

[44] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 2071, 2072, 2076, 2077, 2078