L2G：一个用于弱监督语义分割的局部到全局的知识迁移框架*

Peng-Tao Jiang[1] ✉ Yuqi Yang[1] Qibin Hou[1†✉] Yunchao Wei[2]
[1] TMCC, CS, Nankai University [2] Beijing Jiaotong University

摘要

挖掘精确的类别感知注意力图（又称类激活图），对于弱监督语义分割十分重要。在本文中，我们提出了 L2G，这是一种简单的、从局部到全局的知识迁移框架，可用于高质量地发掘对目标物体的注意力。我们观察到，当用局部的图像块代替整张输入图像时，分类模型能够更够获得更加细致的目标区域。考虑到这一点，我们首先将输入图像随机切成多个局部块，并且利用局部分类网络抽取它们的特征；接着，我们利用全局网络，从多个局部注意力图中在线学习互补的注意力知识。我们的框架引导全局网络学会从全局视角捕获丰富的物体细节知识，进而可以直接使用高质量注意力图作为语义分割网络的伪标注。实验表明，我们的方法在 PASCAL VOC 2012 和 MS COCO 2014 的验证集上分别获得了 72.1% 和 44.2% 的 mIoU，均创造了新的记录。代码存放在 https://github.com/PengtaoJiang/L2G 。

1. 引言

近年来，深度学习算法 [42, 38, 65]推动了语义分割任务的快速发展。然而，为语义分割训练深度神经网络需要大量像素级的精确标注，这十分消耗人力物力。最近，研究者们为了减少对精确标注的依赖，尝试使用一些比较廉价的标注来监督学习语义分割任务，如边界框 [43, 12]、涂鸦 [37, 51]、点 [4]、图像级标签 [54, 23]等。在这些弱监督策略中，图像级标签是最流行的，因为它仅仅需要提供目标物体类别是否存在的信息，容易收集。本文中，我们主要关注基于图像级标签的弱监督语义分割 (WSSS)。

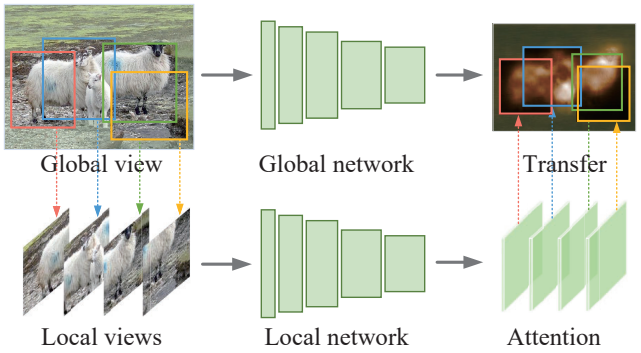

图 1. 所提方法的概念性工作流程我们从局部网络中抽取细节丰富的局部视图的注意力图，并用其教授全局网络，使全局网络在线地从局部网络中学习到丰富的局部细节知识，从而得到更加完整的目标物体注意力。

提到弱监督语义分割，最重要的就是类激活图 (CAM) [66] ，图中包含了语义和位置信息，并且能够用作训练分割网络的伪像素级标注。由于类激活图的质量对于分割的结果有很大影响，最近提出了很多改进类激活图的策略，包括对抗擦除 [54, 64, 22, 67]，在线注意力累积 [26, 25]，种子区域扩张 [29, 24]，相似性学习 [2, 1, 58]等等。尽管这些方法获得了好的性能，但是它们都将整张输入图片作为模型的唯一输入。然而，我们通过实验观察到，当使用局部图像块而非整张图片作为输入时，分类模型能够发现更多可分的区域，这提出了一种利用局部图像块来提高注意力图的质量的适当方法。

考虑上述分析，我们在本文中提出了一种简单的从局部到全局的知识迁移框架，称为 L2G，来生

*本文为 CVPR'22 论文 [27]的中文翻译版。
†Qibin Hou 是通讯作者

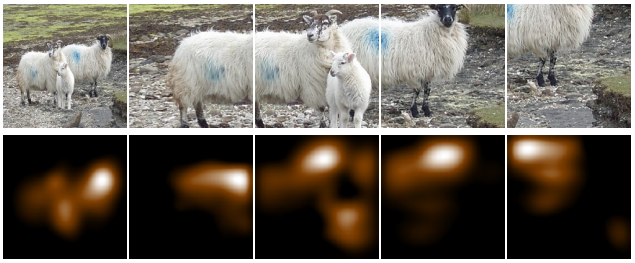

图 2. L2G 注意力知识迁移方法的动机。上一行展示了原始图片（全局视角）和经过随即剪裁的多个图像块（局部视角），第二行展示了通过 CAM[66]生成的对应的注意力图。我们可以观察到，相比于全局视图，局部视图下的注意力图能够捕获更多的物体细节信息。

成高质量的目标物体的注意力。图1中描绘了概念图。与先前的注意力挖掘策略不同，我们提出同时利用输入图像的全局视图和从中随机剪裁的局部视图（被彩色边界框包围的区域）。具体来说，我们的框架包含一个局部网络和一个全局网络，前者为局部视图产生有丰富目标物体细节的局部注意力，后者接受全局视图作为输入，旨在从局部网络中蒸馏出可区分的注意力知识。

我们的方法有如下优点：首先，从输入图像的多个局部视角而非全局视角产生注意力图，使之获得了更多关于未发现的语义区域的细节，且这些区域在不同的局部视图中是互补的，如图2所示；其次，通过设计知识迁移损失，可以将互补的注意力知识通过在线学习的方式高效地迁移到全局网络中，这使得全局网络捕获了像素级别的目标物体细节并在推理时输出高质量的注意力图；最后，整个流程简单而灵活，我们可以有选择地向局部网络添加额外的约束 [33] 来帮助塑造所获得的目标物体注意力。

我们在 PASCAL VOC 2012 和 MS COCO 2014 数据集上评估了我们的方法，实验表明，该方法比以往最先进的方法获得了更好的性能。使用 DeepLab-v2 模型 [10]作为我们的分割网络时，我们在 PASCAL VOC 2012 的验证集和测试集上分别获得了 72.1% 和 71.7% 的 mIoU，在 MS COCO 2014 的验证集上获得了 44.2% 的 mIoU，创造了弱监督预语义分割的新纪录。我们还进行了一系列的消融实验，以帮助读者更好地了解每个组件在我们的方法中的表现。

2. 相关工作

2.1. 弱监督语义分割

单阶段弱监督语义分割. 这种方式直接利用图像级标签来监督训练端到端的分割网络。先前的工作 [44, 45] 将该问题描述为多实例学习。后来，Papandreou 等人 [43] 提出了一种用中间层的预测对分割网络进行监督的期望最大化 (EM) 方法；Zhang 等人 [63]利用图像分类分支生成注意力图，并构造伪分割标注监督与之并行的分割分支；Araslanovet 等人 [3]提出了一种在训练过程中利用图像外观先验知识生成伪分割标注的自监督机制；Chen 等人 [7]构造了一个使用编码器-解码器探索物体边界的端到端框架。与两阶段弱监督语义分割相比，单阶段的方法往往性能较差，吸引力较小。

两阶段弱监督语义分割. 这种方式依靠注意力图 [66, 64] 生成伪分割标注，然后用其训练分割网络。因此，两阶段方法的核心是要得到高质量的注意力图 [55, 5, 53, 49, 34]。为了实现这个目标，最近提出了许多工作。Wei 等人 [54]提出了对抗擦除策略，通过在迭代过程中依次遮挡挖掘出的目标区域来驱动分类网络发现新的目标区域；Hou 等人 [22]用自擦除策略改进了对抗擦除策略，以防止注意力扩散到背景；Kolesnikov 等人 [29]引入了种子扩展的思想，从预先计算的注意力图中扩展出初始种子区域，并约束扩展区域与对象边界对齐；后来，Jiang 等人 [26]提出了利用不同训练阶段的注意力图的在线注意力积累策略；Chang 等人 [6]使用子类别信息突出不可区分的语义区域。

另一系列工作试图细化注意力图，以获得具有宝贵边界的完整的目标物体区域：Ahn 等人 [2]学习了像素相似性，将注意力图中的强响应传播到相邻像素；Chen 等人 [8]和 Ahn 等人 [1]后来使用显式的类边界学习改进了这种方法；Lee 等人 [33]利用现成的显著性图作为监督，指导区域学习生成高质量地注意力图。

上述方法都在全局视图上细化了注意力图。我们的方法与之不同，既利用了全局视图，又利用了

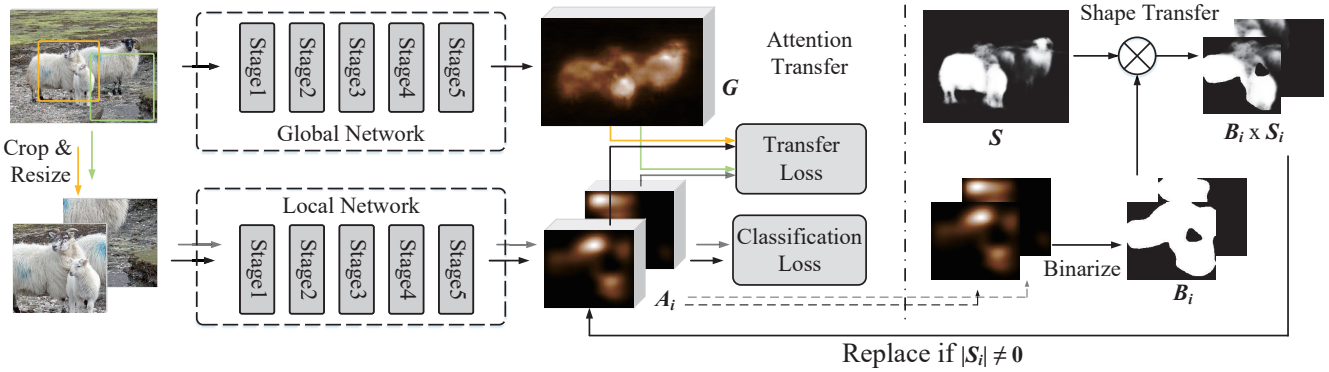

图 3. 所提出方法的整体框架。局部网络捕获的互补的注意力图通过知识迁移损失被蒸馏到全局网络中。

多个局部视图，并且学习了将局部网络中互补的注意力信息高效地迁移到全局网络中的方式，提升了注意力图的质量。

2.2. 知识蒸馏

我们的工作同样与知识蒸馏 [21, 17, 62]有关。知识蒸馏是将训练得很好的教师模型中的知识蒸馏到学生模型中。在图像分类任务中，蒸馏侧重于模仿教师模型的预测分布来改进学生模型。另外，一些研究者 [40, 20]同样研究了对于语义分割的知识蒸馏模型。与之不同的，我们研究了如何以在线学习的方式将本地视图捕获的注意力知识转移到全局网络，以更好地利用来自多个视图的互补信息。

3. 方法

3.1. 先决条件

我们首先介绍生成注意力图的方法，给定输入图像 I，图像级标签 y，最后一层卷积的输出特征是 F 的通道数是 C，等于类别数。最后一个卷积层之后是全局平均池化层，将特征 F 池化为长度为 C 的向量 f_c。我们使用 sigmoid 交叉熵损失函数计算分类损失，其形式如下：

$$L_{\text{ce}} = -\frac{1}{C}\sum_{c=1}^{C} y^c \log(\sigma(f^c)) + (1-y^c)\log(1-\sigma(f^c)) \tag{1}$$

其中，σ 是 sigmoid 函数。注意力图能够从最后一个卷积层的输出生成。对于某个类别 c，注意力图 A^c 是从 F 的第 c 个通道得到的，其形式如下：

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

正如先前的工作 [54, 22, 26, 2]指出的那样，上述方法仅能定位区分度高的区域，而对语义分割同样重要的难以区分的目标物体区域，该方法常常失效。接下来，我们提出了一种新的注意力生成框架时，通过一种新的从局部到全局的知识迁移方法来捕获高质量地对象注意力。

3.2. 整体框架

正如在第1节所提到的，侧重于处理局部图像块的局部网络倾向于发现更多可区分的目标物体区域。基于这样的观察，我们提出借助局部视图的注意力图来帮助全局网络定位更加完整的目标区域。

我们提出的方法的整体框架如图3所示。功能上可以分为四个组件：全局网络、局部网络、注意力迁移模块和形状迁移模块。全局网络和局部网络可以是任何的卷积神经网络分类器，例如流行的 VGGNet [47]或者 ResNet-38 [57]。在注意力迁移模块中，我们优化两个损失函数：一个用于识别分割目标物体的分类损失 L_{cls}，和一个鼓励全局网络模仿局部网络发现更多可区分区域的注意力迁移损失 L_{at}。在形状迁移模块，为了对捕获的目标物体注意力进行塑形，我们向损失 L_{at} 中引入一个形状限制，得到 L_{st}。因此整个损失函数有如下形式：

$$L = L_{\text{cls}} + \lambda \cdot L_{\text{kt}} \tag{3}$$

其中 λ 是损失 L_{kt} 的权重；在没有形状限制时，$L_{\text{kt}} = L_{\text{at}}$，否则 $L_{\text{kt}} = L_{\text{st}}$。

3.3. 局部到全局的注意力迁移

给定输入图像 I，我们将其转化为一系列的不同视图 V，包括全局视图 V_I 和 N 张从全局视图中随机剪裁的局部视图 $\{V_1, V_2, \ldots, V_N\}$。局部视图 $\{V_1, V_2, \ldots, V_N\}$ 被送入局部网络来生成包含丰富目标物体细节的注意力图；全局视图被送入全局网络，旨在从局部网络中学习知识并在推理过程中产生目标物体注意力。记局部网络的最后一个卷积层的输出为 $\{F_1, F_2, \ldots, F_N\}$，每一项都有等同于类别数量的通道数 C。记 $\hat{F}$ 为全局网络最后一个卷积层的输出，通道数为 $C+1$。下面将定义分类损失和注意力迁移损失：

分类损失. 在局部网络上配置了分类损失。具体地说，局部视图的特征图 $\{F_1, F_2, \ldots, F_N\}$ 首先被送入全局池化层，从而得到一个 1D 特征向量的集合 $\{f_1, f_2, \ldots, f_N\}$；给定一个 1D 特征 f_i，对所有类别的概率预测为 $q_i = \sigma(f_i)$，其中 σ 是 sigmoid 函数。这样我们可以将分类损失 L_{cls} 表示为：

$$L_{\text{cls}} = -\frac{1}{N \times C} \sum_{i=1}^{N} \sum_{c=1}^{C} y^c \log(q_i^c) + (1 - y^c) \log(1 - q_i^c) \tag{4}$$

注意力迁移损失. 我们首先从局部网络中生成局部视图下的注意力图。如果 c 在图像级标签中，我们按照式2来生成注意力图 $\{A_1^c, A_2^c, \ldots, A_N^c\}$；如果 c 不在图像级标签中，则对应的注意力图被置为零。为了将局部网络获得的注意力迁移到全局网络中，我们采用均方差损失。

给定全局网格的输出 $\hat{F}$，我们对 $\hat{F}$ 的每个位置沿着通道维度使用 Softmax 函数，得到

$$G^c = \frac{e^{\hat{F}^c}}{\sum_{i=1}^{C+1} e^{\hat{F}_i}} \tag{5}$$

其中 G^c 的每个位置的值意味着该位置属于类别 c 的概率。用 $\{G_1, G_2, \ldots, G_N\}$ 表示与全局视图上 $\{A_1, A_2, \ldots, A_N\}$ 对应的区域，即，每一对 (G_1, A_1) 剪裁自全局视图上的相同位置。注意力迁移损失是通过度量 $\{A_i\}$ 和 $\{G_i\}$ 的差异来制定的：

$$L_{\text{at}} = \frac{1}{N} \sum_{i=1}^{N} ||A_i - G_i||^2 \tag{6}$$

在训练过程中，我们同时优化上述两个损失函数；在推理过程中，注意力图直接由全剧网络生成，局部网络可以抛弃。

讨论. 我们提供了一种高效的方法，它能利用来自全局视图和局部视图的互补信息。这种局部到全局到注意力迁移方法通过在线学习的方式引导全局网络吸收局部网络捕获的丰富的目标物体的细节知识。尽管大多数先前的工作使用了对输入的数据增强，例如随机剪裁，但是它们没有一个组件，利用从全局视图中剪裁的局部图像块，来积累目标物体细节知识；这使得我们的局部到全局的方法与先前的工作有很大的不同。我们将在实验章节展示我们的方法相比于其他方法的更多优点。

3.4. 局部到全局的形状迁移

上述提出的局部到全局的注意力迁移策略已经能够得到一个相比原始类激活图 [66]更加完整的目标物体注意力。然而，作为由于注意力迁移过程只利用了图像级别的标签，捕获的对象边界周围的注意力不够清晰。为了更好地捕捉局部物体在注意力图中的形状，我们尝试在注意力转移损失中通过增加一个形状约束来引入辅助的显著物体信息。显著模型 [39]可以用作类别不可知的显著目标检测器，可以分割前景对象并提供形状信息。

如图3右侧所示，形状迁移过程非常简单。给定局部网络得到的注意力图 $\{A_i\}$，我们首先使用一个小的阈值将其二值化，得到二值图 $\{B_i\}$；接着，我们使用显著性模型生成输入图像 I 的显著性图 S，并将 I 上同一坐标位置下的注意力 $\{A_i\}$ 对应的显著性图记为 $\{S_i\}$。这样，注意力转移损失可以改写为：

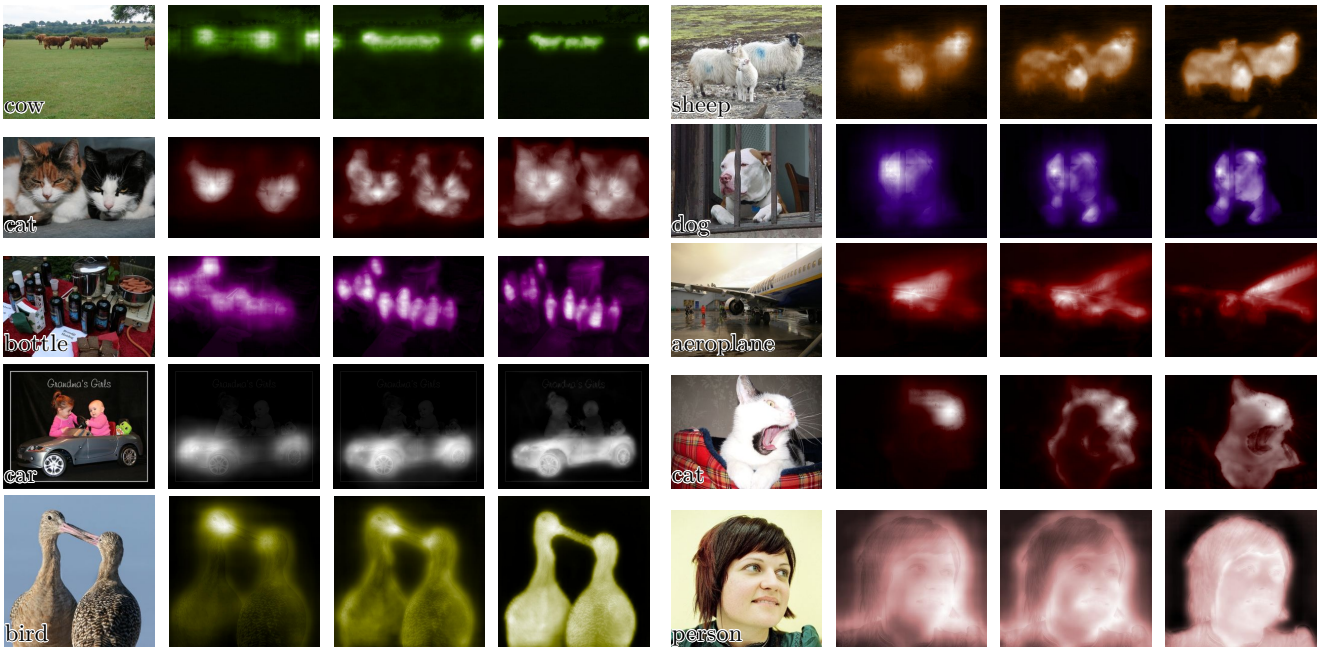

图 4. 来自不同方法的注意力图的定性比较。

$$L_{\text{st}} = \frac{1}{N}\sum_{i=1}^{N}\begin{cases} ||B_i \times S_i - G_i||^2, & \text{if } |S_i| \neq 0 \\ ||A_i - G_i||^2, & \text{if } |S_i| = 0 \end{cases} \tag{7}$$

其中，$\times$ 是逐元素乘法，而 $|S_i|$ 是显著性 S_i 的基数。我们使用 $B_i \times S_i$ 来去除非显著区域的注意力，这些区域通常很大概率属于背景。这允许我们的方法能够充分利用显著图所提供的形状信息，并产生高质量的注意力图。我们将在实验部分详细说明这一点。注意到不是所有图像都有显著目标，总是使用式7上面的一项并不合适，因此对于那些显著性图中什么都没有的图像，我们使用原来的注意力图作为监督，如式7下面一项所示。

值得一提的是，EPS [33]同样使用显著性图作为监督，为网络提供形状信息。与之不同的是，我们的方法关注于如何利用多种局部视图的优点、如何高效地将局部网络学习到的知识迁移到全局网络。在后文中，我们将展示局部到全局的知识迁移相比于 EPS 的优越性。

4. 实验

本文接下来按照如下方式组织：首先，我们介绍了实验的配置；接着，我们展示了消融研究的实验结果，并分析了在我们的方法中提出的每个分量的作用；最后，我们进行了实验，将我们的方法与以前最先进的弱监督语义分割方法进行了比较。

4.1. 实验设置

数据集. 实验在两个公开数据集上进行：PASCAL VOC 2012 和 MS COCO 2014。PASCAL VOC 2012 数据集包含了除背景之外的 20 个语义类别，它被分为训练集、验证集和测试集，分别包含 1464，1449 和 1456 张图片。与先前工作一致地，我们也使用了扩充的训练集 [18]，一共有 10582 副训练图像片。MS COCO 2014 数据集包含了 80 个语义类别，与先前工作 [11, 33]一致地，没有目标类别的图片将从数据集中剔除，留下 82081 副训练图片和 40137 副验证图片。

评估指标. 使用平均交并比 (mIoU) [42] 作为度量指标。由于 PASCAL VOC 2012 数据集的测试集不

表 1. 不同网络设置下 mIoU 分数的比较。基线是原始 CAM [66]；SW: 在推理阶段将滑动窗口策略应用于基线 [66]；Local: 使用多幅局部图像块而非整张输入图像来训练分类网络；L2G: 我们的方法，但仅仅使用局部到全局的注意力迁移；$\text{mIoU}_{trainaug}$ 表示在扩展训练集上伪分割标注的 mIoU 分数。

No.	SW	Local	L2G	$\text{mIoU}_{trainaug}$	mIoU_{val}
1				47.1	47.5
2	✓			46.1 (-1.0)	47.2 (-0.3)
3		✓		48.5 (+1.4)	50.0 (+2.5)
4			✓	56.8 (+9.7)	54.9 (+7.4)

表 2. 关于每个组件重要性的消融实验。L2G: 仅仅使用局部到全局的注意力迁移；Shape: 局部到全局的形状迁移。可以看出。与仅仅是用局部网络的配置相比，我们的局部到全局的迁移策略可以显著提高性能。当结合形状信息时，L2G 仍可以大幅度提升性能。

No.	Local	L2G	Shape	$\text{mIoU}_{trainaug}$	mIoU_{val}
1	✓			48.5	50.0
2		✓		56.8 (+8.3)	54.9 (+4.9)
3	✓		✓	68.0	69.9
4		✓	✓	70.3 (+2.3)	72.1 (+2.2)

可获得，我们将评估结果提交到 PASCAL VOC 官方的评估服务器上[1]。

数据增强. 为了增强数据，输入图像的短边被调整到 512，从中随机剪裁 448×448 作为全局视图。局部视图是从全局视图中随机剪裁的 320×320 图像块。

分类网络. 与先前工作 [2, 33]一致地，我们使用 ResNet-38[57]作为我们的分类网络。另外，我们同样在分类网络中使用一个像素相关模块 (PCM)[53]来限制目标物体的形状。使用多尺度测试策略 [2]从全局网络生成注意力图。

在 PASCAL VOC 上分类. 我们使用 SGD 作为优化器并训练分类网络 10 个周期。初始的学习率设置为 1e-3，从第 6 周期开始衰减。注意力迁移损失的权重 λ 设置为 10。其他网络参数如下：批量大小为 3；权重衰减为 5e-4；图像块尺寸为 320×320；图像块数量为 4。

分割. 我们使用 DeepLab-v1 [9]和 DeepLab-v2[10]作为我们的分割网络。我们同时报告了基于 VGG-16[47] 和基于 ResNet-101[19]的性能。对于基于 VGG-16 的分割网络，我们使用在 ImageNet[13] 上预训练模型进行初始化。对于 MS COCO 数据集上的实验，我们均使用 ImageNet 预训练的模型。我们使用与先前工作 [33]一致地方式生成伪标签。给定注意力图，我们给背景通道分配一个固定的阈值，并使用 argmax 函数来获得每个像素的标签。

4.2. 消融研究

我们设计了多种消融实验来对我们的方法进行健全的检查。所有的消融实验在 PASCAL VOC 2012 数据集上进行。我们报告了在扩展训练集和验证集上分割结果的 mIoU。

局部视图采样策略. 首先，我们研究了采样策略对于注意力图的影响。我们比较了两种局部视图的采样策略：一种是随机采样，另一种是均匀采样。我们使用在全局视图上的滑动窗口实现均匀采样，在这种方式下，每一个像素都可以被包含在某个局部视图中。对于 448×448 的全局视图，我们使用 320×320 的窗口大小和 64 的步长，得到 9 个局部视图。为了与均匀采样有公平的比较，我们在随机采样策略中同样采样 9 个图像块。使用两种策略下伪标签得到的分割质量十分接近（随机策略为 68.8%，均匀策略为 68.5%）。为了灵活地调整局部视图的数量 N，我们选择了随机采样策略。

图像块尺寸与数量 N. 块大小控制着局部视图的空间尺寸，块数量 N 决定了送入局部网络的局部视图数量。为了研究它们对于注意力图质量的影响，我们选择了 5 种不同的块尺寸：$[240 \times 240, 280 \times 280, 320 \times 320, 360 \times 360, 400 \times 400]$；在研究块数量时，我们从 $[1, 2, 4, 8, 16]$ 中选择。正如图5所示，我

[1] http://host.robots.ox.ac.uk:8080/

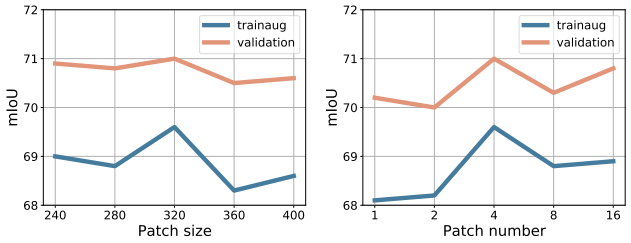

图 5. 关于局部视图尺寸和数量 N 的消融研究。

们观察到当 N 增加时，伪标签的质量逐渐变好，当局部块数量大于 4 时，算法性能趋于稳定。对于图像块尺寸，我们发现在设置为 320×320 时方法获得了最好的性能，当图像块尺寸超过 320×320 时，伪分割标签的质量开始下降。

局部到全局知识迁移的重要性. 当把局部视图送入局部网络时,我们可以从得到的注意力图中发现更多的对象区域。在这里，人们可能会提出一个问题："局部网络生成注意力图是否已经足够好了，不再需要迁移处理？" 为了回答这个问题，我们使用来自局部网络的注意力图来测试伪分割标签的质量。正如表1所示，我们可以看到局部网络的性能 (48.5%) 稍好于 CAM 基线，但远远不如 L2G (56.8%)。我们同样在图4 中定性地展示了一些注意力图的结果，并在图6中展示了一些分割结果。这表明局部到全局的注意力迁移策略是一种更有效地利用局部网络获取丰富对象注意力知识的方式。

此外，通过引入形状信息，我们进一步扩展了上述实验。相应的结果在表2中给出。尽管在 [33]中已经证明，显著形状信息可以提高注意力质量，然而，当使用建议的 L2G 时，在训练集和验证集上的 mIoU 得分能够显著提升。我们将在下一小结展示更多分割的数值结果。

L2G 与滑动窗口对比. 我们的方法的关键是使用局部注意力图来帮助全局网络发现更完整的目标物体区域，一个对该想法直接实现是在推理过程中使用滑动窗口策略，并聚和来自不同图像块的注意力图。我们将 L2G 与滑动窗口策略进行了比较，具体地说，对于滑动窗口策略，我们将窗口大小和窗口数量分别设置为 320×320 和 64；对于 L2G，我们将局部窗口大小设置为 320×320，并一次随机采样 4 个块。

表 3. 不使用显著性图的方法在 PASCAL VOC 训练集上的伪分割标注的比较。

Methods	mIoU_{train}
CAM [66]	48.0
SC-CAM [6]	50.9
SEAM [53]	55.4
ADvCAM [32]	55.6
L2G (ours)	56.2

表 4. 使用了显著性图的方法在 PASCAL VOC 训练集上的伪分割标注的比较。

Methods	mIoU_{train}
SGAN [59]	62.8
EPS [33]	69.4
L2G (ours)	71.9

如表1所示，局部到全局的注意力迁移策略获得了远好于基线 CAM[66]的结果，证明了我们的方法的有效性。然而，滑动窗口策略甚至比原始的 CAM 更糟糕。我们认为，滑动窗口策略不适合挖掘不可分对象区域，因为模型的训练仍然是基于全局视角的输入。这使得在处理全局视图时，在不同区域具有不同外观的未发现物体将难以响应。

全局网络的分类损失. 局部网络配备分类损失来指导注意力图生成。有人可能会问，"全局网络是否需要分类损失?" 为了回答这个问题，我们试着为全局网络添加了分类损失。我们观察到当添加分类损失时，由全局网络生成的注意力图定位到非常小的目标对象区域。伪标注的质量从 70.3% 骤降至 53.8%。我们认为，分类损失与注意力迁移损失起着相反的作用：分类损失让注意力图变得更加可区分，注意力迁移损失有助于将不可取分区域的注意力转移到全局网络。因此，为全局网络添加分类损失后注意力图变得更糟。

局部网络与全局网络共享骨干. 这里，我们探索了是否使用局部网络与全局网络的骨干共享之间的性能差距。当局部和全局网络共享骨干网络时，在扩充训练集上伪分割标注的 mIoU 的得分为 69.2%，在完成分割网络的训练后，在验证集上的 mIoU 为 70.9%。当局部网络与全局网络使用不同的骨干网络时，伪分割标注的得分能够提升 1.1%，最终的分割结果能够获得 1.2% 的 mIoU 的提升。

图 6. 不同网络分割结果的比较。我们可以看到，结合 L2G 和形状迁移的方法得到了最好的结果，尤其体现在局部物体的细节上。

4.3. 与最先进的方法的比较

我们首先将我们生成的注意力图与先前的最先进的弱监督语义分割方法进行比较。我们的注意力图被转换成伪分割标注。如表3和表4所示，无论是否使用显著性图，我们的方法生成的注意力图都要明显好于其他方法。不适用显著性图，我们在 PASCAL VOC 训练集上获得了 56.2% 的 mIoU 得分，比 SEAM [53]高 0.8%。在将显著性图应用到迁移过程中后，mIoU 得分达到了 71.9%，高于 EPS[33]的 69.4%。

我们使用分割标签直接训练 DeepLab 分割模型，并与最先进的方法比较分割性能。表5和表6展示了我们的方法和最近最先进的方法在 PASCAL VOC 数据集上的分割结果，可以看到，与先前的弱监督语义分割方法对比，我们的方法在验证集和测试集上都获得了最好的结果。与我们的工作最相关的是显式使用显著性图进行监督的 EPS[33]。我们的方法与 EPS 的区别在第3.4节中进行了解释。如表6所示，我们的方法相比于 EPS 提升了大约 1%。此外，如表7所示，我们的方法在具有挑战性的 MS

表 5. 在 PASCAL VOC 2012 验证和测试集上与先前最先进的方法进行定量比较。所有的分割结果基于使用 VGGNet [47]作为骨干网络的 DeepLab。Pub.: 出版物, Seg.: 分割网络, Sup.: 监督信号, I.: 图像级标签, S.: 现成的显著性模型生成的显著性图。

Methods	Pub.	Seg.	Sup.	Val (%)	Test (%)
AffinityNet [2]	CVPR'18	V1	I.	58.4	60.5
MCOF [52]	CVPR'18	V1	I.+S.	56.2	57.6
DSRG [24]	CVPR'18	V2	I.+S.	59.0	60.4
SeeNet [22]	NeurIPS'18	V1	I.+S.	61.1	60.7
FickleNet [31]	CVPR'19	V2	I.+S.	61.2	61.9
OAA$^+$ [26]	ICCV'19	V1	I.+S.	63.1	62.8
BES [8]	ECCV'20	V1	I.	60.1	61.1
MCIS [49]	ECCV'20	V1	I.+S.	63.5	63.6
Multi-Est. [16]	ECCV'20	V1	I.+S.	64.6	64.2
ICD [15]	CVPR'20	V1	I.+S.	64.0	63.9
ECS-Net [50]	ICCV'21	V1	I.	62.1	63.4
DRS [28]	AAAI'21	V1	I.+S.	63.5	64.5
Group-WSSS [35]	AAAI'21	V2	I.+S.	63.3	63.6
OAA++$^+$ [25]	PAMI'21	V1	I.+S.	63.7	63.2
NSROM [60]	CVPR'21	V2	I.+S.	65.5	65.3
EPS [33]	CVPR'21	V1	I.+S.	66.6	67.9
EPS [33]	CVPR'21	V2	I.+S.	67.0	67.3
L2G (ours)	–	V1	I.+S.	68.1	68.8
L2G (ours)	–	V2	I.+S.	68.5	68.9

COCO 数据集上远好于先前的方法，这也证明了我们局部到全局策略的有效性。我们的方法得到的伪标注的 mIoU 为 43.4%，远好于 EPS 的 37.2%。

讨论. 值得注意的是，我们的局部网络只是简单的分类网络。提出的框架是灵活的，我们可以将局部网络换位复杂的注意力模型来进一步提高性能。因此，我们相信该框架还有很大的提升空间，我们也希望这种局部到全局的知识迁移方法能够为研究者们提供一个新的方向。

失败样例分析. 首先，一些非目标物体被识别为目标物体，如图7的前两行所示。在 L2G 中，我们仅仅使用 ResNet 来抽取特征。设计更加先进的分类模型，例如 transformers [14, 61]，能够改善结果。其次，被发现目标物体的形状还有很大改善空间（如图7最后两行所示）。使用更强的显著性模型，或者过度分割方法也许能在一定程度上解决该问题。

表 6. 与先前最先进方法在 PASCAL VOC 2012 的验证集和测试集上的定量比较，所有的分割结果都基于 ResNet 骨干网络 [19, 57]，我们的方法获得了最好的结果。

Methods	Publication	Seg.	Sup.	Val (%)	Test (%)
AffinityNet [2]	CVPR'18	V1	I.	61.7	63.7
MCOF [52]	CVPR'18	V1	I.+S.	60.3	61.2
DSRG [24]	CVPR'18	V2	I.+S.	61.4	63.2
SeeNet [22]	NeurIPS'18	V1	I.+S.	63.1	62.8
IRNet [1]	CVPR'19	V1	I.	63.5	64.8
FickleNet [31]	CVPR'19	V2	I.+S.	64.9	65.3
OAA$^+$ [26]	ICCV'19	V1	I.+S.	65.2	66.4
SSDD [46]	ICCV'19	V1	I.	66.1	66.8
SEAM [53]	CVPR'20	V2	I.	64.5	65.7
SC-CAM [6]	CVPR'20	V2	I.	66.1	65.9
ICD [15]	CVPR'20	V1	I.+S.	67.8	68.0
BES [8]	ECCV'20	V2	I.	65.7	66.6
MCIS [49]	ECCV'20	V1	I.+S.	66.2	66.9
Multi-Est. [16]	ECCV'20	V1	I.+S.	67.2	66.7
LIID [41]	PAMI'20	V2	I.+IS.	66.5	67.5
DRS [28]	AAAI'21	V2	I.+S.	71.2	71.4
Group-WSSS [35]	AAAI'21	V2	I.+S.	68.2	68.5
ECS-Net [50]	ICCV'21	V1	I.	66.6	67.6
PMM [36]	ICCV'21	PSP	I.	68.5	69.0
CDA [48]	ICCV'21	V2	I.	66.1	66.8
CGNet [30]	ICCV'21	V1	I.	68.4	68.2
AuxSegNet [58]	ICCV'21	V1	I.+S.	69.0	68.6
AdvCAM [32]	CVPR'21	V2	I.	68.1	68.0
NSROM [60]	CVPR'21	V2	I.+S.	70.4	70.2
EDAM [56]	CVPR'21	V1	I.+S.	70.9	70.6
EPS [33]	CVPR'21	V1	I.+S.	71.0	71.8
EPS [33]	CVPR'21	V2	I.+S.	70.9	70.8
L2G (ours)	–	V1	I.+S.	72.0	73.0
L2G (ours)	–	V2	I.+S.	72.1	71.7

5. 总结

在本文中，我们提出了一个新颖的局部到全局的注意力迁移方法来获得目标物体注意力。利用局

表 7. 与先前最先进方法在 MS COCO 验证集上的定量比较，除了 L2G* 使用 ResNet-101 骨干 [19]，所有的分割结果基于 VGGNet 骨干网络 [47]。

Methods	Publication	Seg.	Sup.	Val (%)
SEC [29]	ECCV'16	V1	I.+S.	22.4
DSRG [24]	CVPR'18	V2	I.+S.	26.0
ADL [11]	PAMI'20	V1	I.+S.	30.8
EPS [33]	CVPR'21	V2	I.+S.	35.7
L2G (ours)	–	V2	I.+S.	42.7
L2G* (ours)	–	V2	I.+S.	44.2

Image Result GT

图 7. 我们的 L2G 的两种失败的分割样例。

部网络从局部视图捕获的互补的注意力，并将形状约束引入注意力转移过程，我们的方法在 PASCAL VOC 2012 的验证集和测试集以及 MS COCO 2014 的验证集上均获得了最好的结果。我们希望提出的方法能够促进依赖高质量注意力图的视觉任务的研究。

参考文献

[1] J. Ahn, S. Cho, and S. Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In IEEE Conf. Comput. Vis. Pattern Recog., pages 2209–2218, 2019. 409, 410, 417

[2] J. Ahn and S. Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 4981–4990, 2018. 409, 410, 411, 414, 417

[3] N. Araslanov and S. Roth. Single-stage semantic segmentation from image labels. In IEEE Conf. Comput. Vis. Pattern Recog., pages 4253–4262, 2020. 410

[4] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei. What' s the point: Semantic segmentation with point supervision. In Eur. Conf. Comput. Vis., pages 549–565, 2016. 409

[5] Y.-T. Chang, Q. Wang, W.-C. Hung, R. Piramuthu, Y.-H. Tsai, and M.-H. Yang. Mixup-cam: Weakly-supervised semantic segmentation via uncertainty regularization. In Brit. Mach. Vis. Conf., 2020. 410

[6] Y.-T. Chang, Q. Wang, W.-C. Hung, R. Piramuthu, Y.-H. Tsai, and M.-H. Yang. Weakly-supervised semantic segmentation via sub-category exploration. In IEEE Conf. Comput. Vis. Pattern Recog., pages 8991–9000, 2020. 410, 415, 417

[7] J. Chen, S. Fang, H. Xie, Z.-J. Zha, Y. Hu, and J. Tan. End-to-end boundary exploration for weakly-supervised semantic segmentation. In ACM Int. Conf. Multimedia, pages 2381–2390, 2021. 410

[8] L. Chen, W. Wu, C. Fu, X. Han, and Y. Zhang. Weakly supervised semantic segmentation with boundary exploration. In Eur. Conf. Comput. Vis., pages 347–362. Springer, 2020. 410, 417

[9] 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 Int. Conf. Learn. Represent., 2015. 414

[10] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell., 40(4):834–848, 2017. 410, 414

[11] J. Choe, S. Lee, and H. Shim. Attention-based dropout layer for weakly supervised single object localization and semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell., 2020. 413, 418

[12] J. Dai, K. He, and J. Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for se-

mantic segmentation. In Int. Conf. Comput. Vis., pages 1635–1643, 2015. 409

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In IEEE Conf. Comput. Vis. Pattern Recog., pages 248–255, 2009. 414

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In Int. Conf. Learn. Represent., 2021. 417

[15] J. Fan, Z. Zhang, C. Song, and T. Tan. Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 4283–4292, 2020. 417

[16] J. Fan, Z. Zhang, and T. Tan. Employing multi-estimations for weakly-supervised semantic segmentation. In Eur. Conf. Comput. Vis., pages 332–348. Springer, 2020. 417

[17] T. Furlanello, Z. Lipton, M. Tschannen, L. Itti, and A. Anandkumar. Born again neural networks. In Int. Conf. Mach. Learn., pages 1607–1616, 2018. 411

[18] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In Int. Conf. Comput. Vis., pages 991–998, 2011. 413

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In IEEE Conf. Comput. Vis. Pattern Recog., pages 770–778, 2016. 414, 417, 418

[20] T. He, C. Shen, Z. Tian, D. Gong, C. Sun, and Y. Yan. Knowledge adaptation for efficient semantic segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 578–587, 2019. 411

[21] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. In Adv. Neural Inform. Process. Syst. Worksh., 2014. 411

[22] Q. Hou, P. Jiang, Y. Wei, and M.-M. Cheng. Self-erasing network for integral object attention. In Advances in Neural Information Processing Systems, volume 31, pages 549–559, 2018. 409, 410, 411, 417

[23] Q. Hou, D. Massiceti, P. K. Dokania, Y. Wei, M.-M. Cheng, and P. H. Torr. Bottom-up top-down cues for weakly-supervised semantic segmentation. In Int. Worksh. on Energy Minimization Methods in Comput. Vis. Pattern Recog., pages 263–277, 2017. 409

[24] Z. Huang, X. Wang, J. Wang, W. Liu, and J. Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In IEEE Conf. Comput. Vis. Pattern Recog., pages 7014–7023, 2018. 409, 417, 418

[25] P.-T. Jiang, L.-H. Han, Q. Hou, M.-M. Cheng, and Y. Wei. Online attention accumulation for weakly supervised semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell., 2021. 409, 417

[26] P.-T. Jiang, Q. Hou, Y. Cao, M.-M. Cheng, Y. Wei, and H.-K. Xiong. Integral object mining via online attention accumulation. In Int. Conf. Comput. Vis., pages 2070–2079, 2019. 409, 410, 411, 417

[27] P.-T. Jiang, Y. Yang, Q. Hou, and Y. Wei. L2g: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 16886–16896, 2022. 409

[28] B. Kim, S. Han, and J. Kim. Discriminative region suppression for weakly-supervised semantic segmentation. In AAAI Conf. Artif. Intell., pages 1754–1761, 2021. 417

[29] A. Kolesnikov and C. H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In Eur. Conf. Comput. Vis., pages 695–711, 2016. 409, 410, 418

[30] H. Kweon, S.-H. Yoon, H. Kim, D. Park, and K.-J. Yoon. Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation. In Int. Conf. Comput. Vis., pages 6994–7003, 2021. 417

[31] J. Lee, E. Kim, S. Lee, J. Lee, and S. Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In IEEE Conf. Comput. Vis. Pattern Recog., pages 5267–5276, 2019. 417

[32] J. Lee, E. Kim, and S. Yoon. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 4071–4080, 2021. 415, 417

[33] S. Lee, M. Lee, J. Lee, and H. Shim. Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In IEEE

Conf. Comput. Vis. Pattern Recog., pages 5495–5505, 2021. 410, 413, 414, 415, 416, 417, 418

[34] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu. Tell me where to look: Guided attention inference network. In IEEE Conf. Comput. Vis. Pattern Recog., pages 9215–9223, 2018. 410

[35] X. Li, T. Zhou, J. Li, Y. Zhou, and Z. Zhang. Group-wise semantic mining for weakly supervised semantic segmentation. In AAAI Conf. Artif. Intell., pages 1984–1992, 2021. 417

[36] Y. Li, Z. Kuang, L. Liu, Y. Chen, and W. Zhang. Pseudo-mask matters in weakly-supervised semantic segmentation. In Int. Conf. Comput. Vis., pages 6964–6973, 2021. 417

[37] D. Lin, J. Dai, J. Jia, K. He, and J. Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 3159–3167, 2016. 409

[38] G. Lin, C. Shen, A. van den Hengel, and I. Reid. Efficient piecewise training of deep structured models for semantic segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., 2016. 409

[39] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang. A simple pooling-based design for real-time salient object detection. In IEEE Conf. Comput. Vis. Pattern Recog., pages 3917–3926, 2019. 412

[40] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang. Structured knowledge distillation for semantic segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 2604–2613, 2019. 411

[41] Y. Liu, Y.-H. Wu, P.-S. Wen, Y.-J. Shi, Y. Qiu, and M.-M. Cheng. Leveraging instance-, image-and dataset-level information for weakly supervised instance segmentation. IEEE Trans. Pattern Anal. Mach. Intell., 2020. 417

[42] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 3431–3440, 2015. 409, 413

[43] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In Int. Conf. Comput. Vis., pages 1742–1750, 2015. 409, 410

[44] D. Pathak, E. Shelhamer, J. Long, and T. Darrell. Fully convolutional multi-class multiple instance learning. In Int. Conf. Learn. Represent., 2015. 410

[45] P. O. Pinheiro and R. Collobert. From image-level to pixel-level labeling with convolutional networks. In IEEE Conf. Comput. Vis. Pattern Recog., pages 1713–1721, 2015. 410

[46] W. Shimoda and K. Yanai. Self-supervised difference detection for weakly-supervised semantic segmentation. In Int. Conf. Comput. Vis., pages 5208–5217, 2019. 417

[47] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In Int. Conf. Learn. Represent., 2015. 411, 414, 417, 418

[48] Y. Su, R. Sun, G. Lin, and Q. Wu. Context decoupling augmentation for weakly supervised semantic segmentation. In Int. Conf. Comput. Vis., 2021. 417

[49] G. Sun, W. Wang, J. Dai, and L. Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In Eur. Conf. Comput. Vis., pages 347–365, 2020. 410, 417

[50] K. Sun, H. Shi, Z. Zhang, and Y. Huang. Ecs-net: Improving weakly supervised semantic segmentation by using connections between class activation maps. In Int. Conf. Comput. Vis., pages 7283–7292, 2021. 417

[51] P. Vernaza and M. Chandraker. Learning random-walk label propagation for weakly-supervised semantic segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 7158–7166, 2017. 409

[52] X. Wang, S. You, X. Li, and H. Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In IEEE Conf. Comput. Vis. Pattern Recog., pages 1354–1362, 2018. 417

[53] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 12275–12284, 2020. 410, 414, 415, 416, 417

[54] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In IEEE Conf. Comput. Vis. Pattern Recog., pages 1568–1576, 2017. 409, 410, 411

[55] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 7268–7277, 2018. 410

[56] T. Wu, J. Huang, G. Gao, X. Wei, X. Wei, X. Luo, and C. H. Liu. Embedded discriminative attention mechanism for weakly supervised semantic segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 16765–16774, 2021. 417

[57] Z. Wu, C. Shen, and A. Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. Pattern Recogn., 90:119–133, 2019. 411, 414, 417

[58] L. Xu, W. Ouyang, M. Bennamoun, F. Boussaid, F. Sohel, and D. Xu. Leveraging auxiliary tasks with affinity learning for weakly supervised semantic segmentation. In Int. Conf. Comput. Vis., pages 6984–6993, 2021. 409, 417

[59] Q. Yao and X. Gong. Saliency guided self-attention network for weakly and semi-supervised semantic segmentation. IEEE Access, 8:14413–14423, 2020. 415

[60] Y. Yao, T. Chen, G.-S. Xie, C. Zhang, F. Shen, Q. Wu, Z. Tang, and J. Zhang. Non-salient region object mining for weakly supervised semantic segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 2623–2632, 2021. 417

[61] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, and S. Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In Int. Conf. Comput. Vis., pages 558–567, 2021. 417

[62] L. Yuan, F. E. Tay, G. Li, T. Wang, and J. Feng. Revisiting knowledge distillation via label smoothing regularization. In IEEE Conf. Comput. Vis. Pattern Recog., pages 3903–3911, 2020. 411

[63] B. Zhang, J. Xiao, Y. Wei, M. Sun, and K. Huang. Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In AAAI Conf. Artif. Intell., pages 12765–12772, 2020. 410

[64] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. S. Huang. Adversarial complementary learning for weakly supervised object localization. In IEEE Conf. Comput. Vis. Pattern Recog., pages 1325–1334, 2018. 409, 410

[65] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In IEEE Conf. Comput. Vis. Pattern Recog., pages 2881–2890, 2017. 409

[66] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In IEEE Conf. Comput. Vis. Pattern Recog., pages 2921–2929, 2016. 409, 410, 412, 414, 415

[67] K. Zhou, Q. Hou, Z. Li, and J. Feng. Multiminer: Object-adaptive region mining for weakly-supervised semantic segmentation. arXiv preprint arXiv:2006.07834, 2020. 409