大规模无监督语义分割

高尚华, 李钟毓, 杨明玄, 程明明, 韩军伟, Philip Torr

摘要—借助大型数据集，例如 ImageNet 和 Microsoft COCO，大规模数据的无监督学习在分类任务方面取得了重大进展。然而，是否能够实现大规模无监督语义分割仍然是未知的。实现这个任务有两个主要的挑战: 1) 我们需要一个评测算法好坏的大规模语义分割基准；2) 我们需要研究新的方法来以无监督的方式同时学习类别和形状表征。在这项工作中，我们提出了一个新的问题，即大规模无监督语义分割 (large-scale unsupervised semantic segmentation，LUSS)，并创建了一个基准数据集来跟踪研究进展。基于 ImageNet 数据集，我们提出了 ImageNet-S 数据集，其中包含 120 万张训练图像和约 5 万张用于评测的高质量语义分割标注。我们的基准具有高度的数据多样性和明确的任务目标。我们还针对 LUSS 任务提出了一种简洁而有效的方法。此外，我们还相应地对相关的无监督、弱监督、有监督方法进行了基准测试，指明了 LUSS 任务的挑战和可能的研究方向。基准和源代码可在以下网站公开获取：https://github.com/LUSSeg.

Index Terms—**大规模**, **无监督**, **语义分割**, **自监督**, ImageNet

1 引言

语义分割任务 [1], [2], [3], [4], [5], [6], 旨在用类别信息来分类图像像素，因其重要性而吸引了广泛的研究关注。由于这项任务的固有挑战，大多数工作集中于在多样性有限 [7], [8], [9] 和数据规模较小 [10], [11] 条件下的语义分割。例如，PASCAL VOC 数据集只包含 2K 张图片，而 BDD100K [9] 数据集只关注道路场景。许多方法在这些受限的环境中取得了显著的结果 [12], [13], [14], [15], [16], [17], [18], [19], [20]。然而大幅地扩大问题规模往往会导致研究模式的改变，例如，从 PASCAL VOC [10] 拓展到 ImageNet [21] 使识别任务难度大幅增加。这促使我们思考一个更具有挑战性的问题：语义分割是否可能用于具有广泛多样性的大规模现实世界环境?

然而，由于巨大的数据规模和隐私问题，人为像素级地甚至是图像级地数据标注是十分昂贵的。缺乏足够的基准数据限制了大规模语义分割任务的发展。当使用数百万张甚至数十亿张图片进行训练，例如 ImageNet [21], JFT-300M [22], and Instagram-1B [23]，分类模型的无监督学习已经展现了和有监督学习相当的能力 [24], [25], [26]。为了实现面向真实世界的语义分割，我们提出了一个新的问题：大规模无监督语义分割 (Large-scale Unsupervised Semantic Segmentation, LUSS). 如图1所示，LUSS 任务的目标是在没有人工标注监督的情况下，为大规模图片数据中的每个像素分配类别标签。为了实现这一目标，需要同时解决例如大规模数据下的形状和类别表征学习以及无监督语义聚类等许多挑战。具体来说，模型需要提取具有类别和形状线索的语义表征。类别相关的表征用来区分不同事物的类别，而例如物体、边缘等形状相关的像素级表征对语义分割至关重要。以上两种表征的有效共存对 LUSS 至关重要，因为冲突的表征可能会导致错误的语义分割结果。从大规模数据中生成类别需要鲁棒和高效的语义聚类算法。为像素指定标签需要区分相关和不相关的语义区域。解决 LUSS 的这些挑战也能帮助许多相关的任务。例如，从 LUSS 中学习到的形状表征可以被用作在数据规模和多样性有限条件下的语义分割 [5], [12] 和实例分割 [27] 等像素级下游任务的预训练表征。此外，利用半监督学习范式来微调 LUSS 模型能够在实际应用中只需人工标注一小部分数据即可实现大规模数据的语义分割。

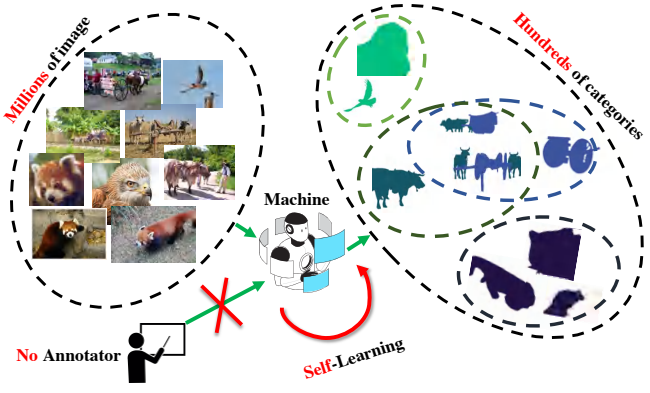


图 1. Large-scale Unsupervised Semantic Segmentation (LUSS) **任务的目标是在没有人类标注的情况下，模型通过自主学习来进行语义分割，将成百上千个类别标签分配给数百万以上的图像中的每个像素。**

为助力 LUSS 任务的发展，我们提出了一个评测基准，其中包含高度多样性的大规模数据，无需直接/间接的人工标注的无监督语义分割任务目标，以及从不同角度进行评测的多种指标。高度多样性的大规模数据给 LUSS 带来了挑战的同时也为模型提供获取丰富的表征的来源。因为数据不足，一些无监督分割方法 [28], [29], [30], [31] 主要关注类别和多样性

- 高尚华、李钟毓、程明明来自南开大学计算机学院。通讯作者程明明邮箱为：cmm@nankai.edu.cn。
- 杨明玄来自加州大学默塞德分校。
- 韩军伟来自西北工业大学。
- Philip Torr 来自牛津大学。

有限的小规模数据的场景，因此不适合 LUSS 任务。基于类别表征学习工作中常用的 ImageNet 数据集 [21]，我们提出了一个用于 LUSS 任务的大规模基准数据集 ImageNet-S。我们移除了 ImageNet 中不可分割的类别，例如书店，并且使用剩余的包含919 个类别的大约 120 万张图片用于训练。然后我们在 ImageNet 验证集中标注了 4 万张图片的精准的像素级语义分割掩码用于 LUSS 任务的评测。我们还标注了训练集中大约 9 千张图像，以支撑更全面的评测和对潜在应用的探索。基于 [32] 中更精确的重新标注的图像级标签，我们使 ImageNet 数据集中的单一图片中有多个类别标签。ImageNet-S 数据集提供的大规模和高多样性的数据能够用于 LUSS 方法的公平训练和充分评测对比。

我们提出了一个用于 LUSS 任务的新方法，包括无监督表征学习，标签生成和微调步骤。对于无监督表征学习，我们提出了 1) 一种非对比像素级表征对齐策略，以在不损害实例级类别表征的情况下增强像素级形状表征；2) 一个提高网络中间层特征表征质量的由深到浅的监督策略。该策略学习到的表征法保证了形状和类别信息的有效共存。我们提出了一种像素注意力方案用来突出有意义的语义区域进而用于像素级标签生成，实现在大规模数据下高效的像素级标签生成和微调。基于提出的方法和 ImageNet-S 数据集，我们分析了本文提出的 LUSS 任务和一些相关工作的关系 (例如无监督学习 [25], [26], [33], [34], [35], 弱监督语义分割 [36], [37], [38], 和下游任务的迁移学习 [5], [27]) 并且确定了 LUSS 任务面临的挑战和可能的研究方向。在这项工作中，我们有两个主要贡献：

- 我们提出了一个新的大规模无监督语义分割问题，以及包含近五万张像素级标注图像、919 个类别和多个评价指标的 ImageNet-S 数据集。
- 我们提出了一种包括增强的表征学习策略和像素注意力机制的新的 LUSS 方法，并且评测了 LUSS 任务的相关工作。

2 相关工作

2.1 无监督分割

在深度学习取得最近的进展之前，很多非参数化方法 (例如标签转移 [39]，匹配 [40], [41]，距离评测 [42]) 和手工设计特征 (例如边缘 [43] 和超像素 [44]) 已被提出用于分割物体。一些无监督分割 (US) 方法只关注分割物体而忽略其类别，然而 LUSS 任务同时关注物体的分割和分类。即便如此，US 模型仍然可以为 LUSS 模型提供先验知识。近期一些数据驱动深度学习模型被用于有监督语义分割任务 [1], [2], [4], [5], [45]。基于现有的预训练表征，[28], [29], [30], [31]，一些无监督语义分割 (USS) 模型使用分割排序 [30]，相互信息最大化 [28]，区域对比学习 [31]，以及几何一致性 [46] 等技术完成该任务。作为 USS 任务的拓展，LUSS 任务不同于 USS 任务的是它的大规模数据和类别。然而，以下几个问题限制了 USS 对 LUSS 任务的适用性：1) 现有模型关注于小数据规模 [28], [29], [30], [31] 和少量 (例如 20+) 且简单的类别 [28](例如天空和地面)。因为数据不足，从大规模数据中无监督学习到的丰富表征的优势没有被现有方法探索。大规模数据所面临的挑战 (例如巨大的计算成本) 也被忽略。2) 由于缺少明确的问题定义和标准的评价指标，一些现有方法使用有监督学习的先验知识，例如有监督预训练网络权重 [46]，有监督边缘检测 [30] 和有监督显著性检测 [31], [47], [48]，使得公平评测这些方法变得困难。

2.2 自监督表征学习

LUSS 任务依赖于自监督学习 (SSL) 提供的语义特征。SSL 方法有助于模型通过代理任务学习语义特征 [49], [50], [51], [52]，例如彩色化 [53], [54], [55]，拼图 [56], [57], [58]，修补 [59]，对抗学习 [60], [61]，上下文预测 [62], [63]，计数 [64]，旋转预测 [58], [65]，跨域预测 [66]，对比学习 [24], [25], [26], [67], [68], [69]，非对比学习 [34], [35], [70]，以及聚类 [33], [71], [72]。我们将介绍与 LUSS 任务相关的几种 SSL 方法。

基于对比学习的 SSL。作为无监督对比学习方法的核心 [67], [73], [74], [75], [76], [77], [78], [79], [80]，基于对比损失的实例区分方法 [81], [82], [83] 利用图像的不同视角 [24], [69] 或者数据增强 [25], [26] 作为正样本对。进而，该类方法迫使模型通过推开负样本对和拉近正样本对来学习表征。Wu 等人 [84] 引入了一个记忆库来扩大可用的负样本进行对比学习。MoCo [25] 用一个动量编码器来稳定训练。CMC [24] 提出了多视角的对比学习，而 SimCLR [26] 探讨了不同数据增强的影响。

基于非对比学习的 SSL。一些非对比学习的方法 [35], [85], [86] 通过最大化图像的不同版本的特征的相似性并避免实用负样本对。BYOL [34] 通过预测由动量编码器输出的特征来避免模型输出崩溃为一个常数的平凡解。SimSiam [70] 利用了梯度停止操作以避免模型训练崩溃。然而，因为对比学习和非对比学习的方法都不包含类别信息，所以它们都在类别相关的任务上表现欠佳，例如，在这些方法中同一类别的实例不一定具有相似的表征。

基于聚类的 SSL。另一个研究方向是将聚类策略引入无监督学习 [87], [88], [89], [90], [91], [92], [93] 来鼓励一组图像具有接近聚类中心的特征表征。Asano 等人 [71] 提出通过一个优化目标来同时进行聚类和表征学习。Li 等人 [72] 通过期望最大化框架最大化观测数据的对数似然，进而迭代执行聚类和对比学习。SwAV [33] 在对视图进行聚类的同时加强聚类簇之间的平衡性。与其他表征学习方法相比，聚类策略有助于通过聚类中心实现更强的类别相关表征。

像素级 SSL。一些工作在像素层级而不是图像层级进行自监

督学习，以增强向像素级下游任务的转移学习能力 [35], [94], [95]。PixPro [35] 在相邻/其他像素之间应用对比学习，并提出像素传播的一致性机制，以增强表征的空间平滑度。SCRL [94] 随机裁剪局部区域，并使与该位置匹配的其他区域具有一致的空间表征。DenseCL [95] 通过匹配图像的两个视图中最相似的特征向量来选择正样本对。尽管在迁移学习方面表现良好，但这些方法忽略了 LUSS 任务所需的实例级别类别相关表征能力。

2.3 弱监督语义分割

弱监督语义分割 (WSSS) [96], [97], [98] 旨在使用例如图像级标签等弱标签完成语义分割任务。由于两者都需要形状表征，WSSS 与 LUSS 有一定相关性相关。然而，在典型 WSSS 方法中的一些设计，例如有监督的 ImageNet_{1k} 预训练模型 [37], [99], [100], [101], 人工标注的图像级标签 [101], [102]，和大参数量的网络架构 [36], [37], 使其不适用于 LUSS 任务。另外，我们仍然可能使用其他 WSSS 中的设计，例如相关性预测 [99], [100], [103]，区域分隔 [102], [104]，边界细化 [99], [105]，联合学习 [106]，和子类别探索 [37]，来改进 LUSS 模型。

3 大规模无监督语义分割基准

LUSS 任务旨在不使用直接/间接人工标注的前提下从大规模图像中学习语义分割。给定大规模图像，LUSS 模型将自学习得到的标签分配给所有图像的每个像素。为了便于理解，我们给出了实现 LUSS 的其中一个方案，见第4节。LUSS 模型同时从大规模数据中学习类别和形状表征，而无需人工标注。该模型使用学习的特征表征进行类别标签聚类和分配，以生成图像的像素级标签。然后，根据生成的标签对模型进行微调，以优化分割结果。理想情况下，标签分配和微调步骤可以隐含在无监督的表征学习过程中。

LUSS 面临多重挑战，例如语义表征学习，大规模数据下的类别标签生成，和无监督学习。此外，缺乏评测基准限制了 LUSS 任务的发展。因此，我们制定了具有明确目标、大规模训练数据和全面评价标准的 LUSS 基准。

3.1 大规模 LUSS 数据集: ImageNet-S

LUSS 任务非常具有挑战性，因为它不使用人工标注标签进行训练，并且需要大规模数据来学习丰富的表征。原则上，LUSS 所需的训练图像规模随着图像复杂性的增长而增加，例如更多的类别数和复杂的场景需要更多的训练数据。现有的分割数据集由于图像复杂度大而数据规模小，很难支持 LUSS 任务。例如 PASCAL VOC [10] 和 CityScapes [7] 等一些数据集仅包含在少数场景下有限数量的图像。例如 ADE20K [8], COCO [107]，和 COCO-Stuff [11] 等其他数据集仅有每个类别的样本数量有限的复杂图像，而对于 LUSS 模型来说很难用有限的数据学习复杂场景的丰富表征。

表 1. ImageNet-S 数据集和现有的语义分割数据集图像类别和数量比较。

Dataset	category	train	val	test
PASCAL VOC 2012 [10]	20	1,464	1,449	1,456
CityScapes [7]	19	2,975	500	1,525
ADE20K [8]	150	20,210	2,000	3,000
ImageNet-S_{50}	50	64,431	752	1,682
ImageNet-S_{300}	300	384,862	4,097	9,088
ImageNet-S	919	1,183,322	12,419	27,423

为了弥补这些数据集的缺陷，常见的有监督分割方法 [5], [12], [27] 使用广泛使用的大型 ImageNet 数据集预训练的模型 [108], [109], [110], [111] 进行微调实现分割。然而，最近的研究 [112], [113] 表明，由于数据分布、数据域和任务目标的不稳定性，ImageNet 和下游数据集上的性能并不总是一致的。对于 LUSS 任务来说，微调预训练的模型使得评价复杂化，并且可能导致不公平和有偏见的比较。ImageNet 有更多的类别、更大的数据规模、相对简单的图像，以及针对每个类别的足够图像，这使得模型学习丰富的表征成为可能。因此，ImageNet 被很多无监督学习方法 [24], [25], [26], [33], [35] 广泛使用。然而，ImageNet 仅有图像级的标注，因此不能用于 LUSS 任务中像素级的评测。为了促进 LUSS 任务，我们从 ImageNet 数据集 [21] 中收集数据并提出了一个大规模 ImageNet-S 数据集并且为 LUSS 评测标注了像素级标签。我们移除了例如书店等不可分割类别，使用了在 ImageNet 中剩余的919 个类别。ImageNet-S 数据集（见图2）比现存的数据集在图像数量（见表1）和类别多样性（见图3）都更大。

3.1.1 图像标注

我们在 ImageNet-S 数据集中标注验证/测试集和部分训练集，以进行 LUSS 评测。因为 ImageNet 数据集有错误的标签并且缺少多类别标签，我们依照 [32] 中重新标注的图像级标签标注像素级语义分割掩码，并进一步更正缺失和错误的标注。图像级标签对应的对象被标注，其他部分被标注为“其他”类别。“其他”类别代表这些类别不经常出现在数据集或只出现在周围环境中。对于验证/测试集，我们标注了919 个所选类别中的所有对象。难以区分的部分标记为“忽略”，不会用于评测。对于训练集，我们对每个类别随机选取十张图像并且标注对应于该类别的对象，而属于919 个类别的其他对象被标注为“忽略”。

语义分割掩码标注。给定一张图像的类别，标注者被要求标注相应的区域并指定正确的类别。在 ImageNet-S 数据集中选定的919 个类别有很高的多样性。一些实例即使是给定类别也不能简单地区分，例如不同品种的狗或不寻常的事物。为了减少标注者识别类别的难度，我们将同一类别的四个图像拼接成一个四方形图像。在这种情况下，标注者可以轻松区分四方形图像中的共有类别。对于具有多个复杂类别的图像，

图 2. ImageNet-S **数据集可视化**.

图 3. ImageNet-S **数据集的类别结构树**。

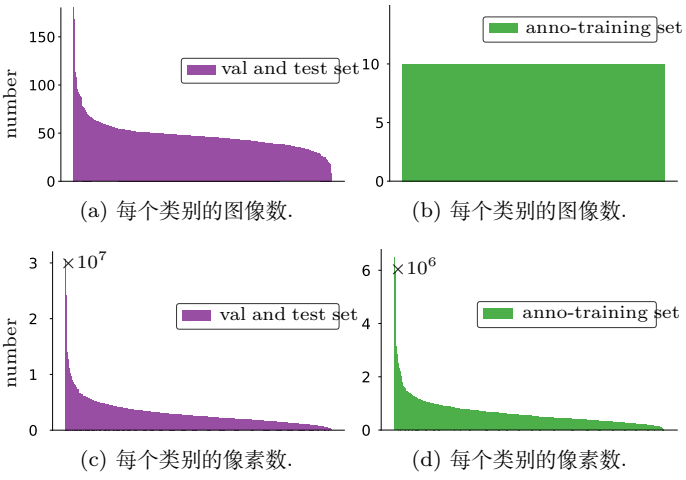


图 4. 在 ImageNet-S **数据集的类别中实例级/像素级数量分布，即每个类别的图像/像素数**。

表 2. ImageNet-S **数据集中每个图像中的类别数**。

	Number of images					
	val set			test set		
Categories in each image	1	2	>2	1	2	>2
ImageNet-S_{50}	745	7	0	1,676	6	0
ImageNet-S_{300}	3,971	118	8	8,815	264	9
ImageNet-S	11,294	954	171	25,133	1,938	352

将提供包含所需类别的多组图像，以帮助标注者识别类别。由于 ImageNet-S 数据集中的类别遵循树状结构（见图3），因此会为不同的标注者提供来自单词树不同子集的图像，以进一步降低标注难度。分辨率低于 1,000 × 1,000 的图像被调整大小到 1,000 × 1,000。标注者将多边形掩码绘制到与类别相关的区域上，每个图像的轮廓上大约有 400 到 500 个点。在调整大小后的高分辨率图像上进行标注可以得到精确的像素级语义分割掩码。

标签标注流程。此数据集的标注团队包括一个组织者，四个质量检查员和15 个标注者。具体的标签标注流程如下：

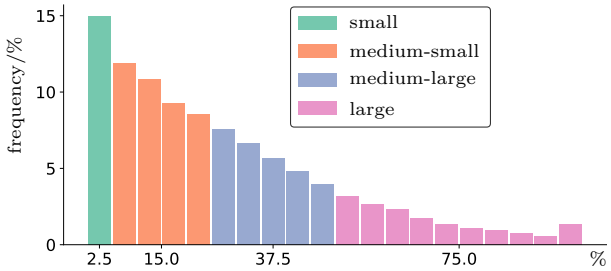
Step 1. 标注者将了解如何标注标签。然后要求标注者按照说明标注一组随机选取的图像。质量检查员检查这些标注好的图像，并纠正错误案例，并将其作为示例展示给所有标注者。

Step 2. 标注者分为几个小组，每个小组有一个小组长。然后组织者将图片分配给每组标注者。标注完图像后，小组长汇总所有的标注并检查标注质量。其他的标注者检查来自组中组长的标注质量。

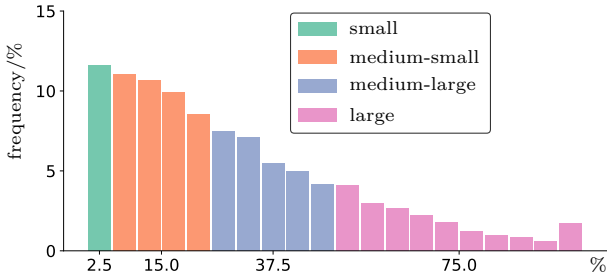
Step 3. 之后将检查过的标注给质量检查员。质量检查员检查标注并给出关于错误案例的反馈。将常见的错误案例和相关的标注解释发送给所有的小组来提高接下来图像的标注质量。

Step 4. 组织者随机检查图像和相应的标注来确保标注质量。

纠正缺失/不正确的标签。在标注过程中，我们发现由于 ImageNet 高多样性和大规模属性，在 [32] 的更正标注中仍有一些缺失和错误的图像级标注。因此，我们提出了几种方案来尽可能地纠正标签：1) 我们发现有些类别是相互关联的，例如蜘蛛和蜘蛛网通常出现在同一张图片中。基于最初标注者观察到的缺失类别，我们对该类别与其他类别相关的图像进行了复查。2) 我们使用例如 Swin transformer [114] 和 Res2Net [110] 等有监督训练图像的分类器，通过检查分类器预测出高置信度但不是 GT 标签的类别，来找到缺失的类别标注。通过这些方案，我们纠正了296 个错误标注的图像并且发现了942 个缺少标签的图像。



(a) 验证集和测试集的分布.



(b) 训练集已标注图像的分布.

图 5. ImageNet-S **数据集中物体大小的分布。物体大小定义为物体与图像大小的比率。**

3.1.2 统计和分布。

图像数量。如表1所示，在 ImageNet 数据集中移除了例如书店，山谷和图书馆这样不可分割类别之后，ImageNet-S 数据集包含919 个类别，1,183,322 张训练图像, 12,419 张验证图像，和27,423 张测试图像。现有的许多自监督表征学习方法 [25], [35] 使用 ImageNet 数据集训练。为了公平比较，我们使用包含1,281,167 张训练图片的 ImageNet 数据集学习无监督表征，并使用 ImageNet-S 数据集进行 LUSS 的其他过程。我们用精确的像素级掩码标注了39,842 张验证/测试图片和9,190 张训练图片，并且在图2可视化了一些标注。我们的像素级标注使得 ImageNet-S 数据集在每张图片中有多个类别。表2 给出了在 ImageNet-S 验证/测试集中每张图片的类别数。大量的图像包含一个类别，8.6% 的图像有多于一个类别。ImageNet-S 相比现有的分割数据集有更简单的图像和更多的类别，它适合于 LUSS 任务下没有人工标注、大规模的图像和大量的类别带来的困难。

类别分布。如图3, ImageNet-S 数据集中的类别由于是从单词树 [21] 中提取的，因此其展现了一个树形结构。图4 展示了 ImageNet-S 数据集类别与图像、像素的数量分布，即每个类别中包含的图像/像素的数量。训练集和验证/测试集有相似的分布。大多数类别的图像数量是均衡的，而每个类别的像素数量呈现长尾分布。像素级类别分布不平衡可能会带来图像级表征学习中未考虑的新挑战。与验证集中每个类别的图像数量相似的原始 ImageNet 数据集相比，重新标记的 ImageNet-S 验证/测试集的图像数量相对于类别而言更不平衡。

表 3. 使用100 个随机选取图像在四个标注者之间的标注一致性。基于 [116]，d 表示像素距离。

Metrics	d	All	S.	M.S.	M.L.	L.
Boundary mIoU [116]	2%	92.4	91.0	91.5	92.7	93.4
	3%	94.8	92.6	93.9	95.1	95.5
	4%	95.9	93.2	95.1	96.2	96.5
Mask mIoU	-	98.7	93.4	97.1	99.0	99.3

物体大小。因为分割更小的物体会更难，我们根据物体与图像的比例将物体分为以下几组，即 small (0%-5%), medium-small (5%-25%), medium-large (25%-50%), and large object size (50%-100%)。图5 中展示的物体大小分布展示出大多数的物体相对较小。

位置分布。我们重叠来自验证和测试集的分割掩码，以分析语义物体在数据集中的位置分布，见图6 (top)。ImageNet-S 数据集中的物体具有中心偏向分布，这说明了现有自我监督方法 [25], [26] 的中心裁切策略的有效性。我们还重叠了物体的边界，见图6 (down)。它表明物体几乎覆盖了所有区域，而不仅仅是图像的中心区域。此外，我们将 ImageNet-S 数据集与 COCO [107] 和 Open Images [115] 数据集的分布进行比较，见图6。ImageNet-S 数据集和其他两个数据集具有相似的分布。所有数据集都观察到了中心偏态分布，我们猜测人类可能倾向于记录更多的中心偏态图像。有趣的是，ImageNet-S 的分布图几乎与 Open Images 数据集相同，而后者以其真实特性而闻名。

标注一致性。我们通过评测不同人的标注一致性来验证标注质量。我们要求四名标注者对 100 张随机选取的图像进行标注。根据四组样本，我们使用掩码 mIoU 和边界 mIoU 评测每对标注之间的平均度量，见表3。掩码 mIoU 达到 98.7%，显示出极高的标注一致性。当 d 为 2% 时, 边界 mIoU 仍有 92.4%, 这表示很高的边缘标注一致性。我们从视觉上观察到，不同人主要的标注差异在边界区域。通过比较不同大小的物体，较小的物体具有较低的标注一致性，因为边界区域在较小物体的标注掩码中占据较大比例。

在有限预算下的 ImageNet-S-50/300。为了在低计算资源预算下促进研究，我们提出了两个包含50 和300 类别的子集，名为 ImageNet-S_{50} and ImageNet-S_{300}。考虑到 LUSS 任务的艰巨性，我们为 ImageNet-S_{50} 50 个在日常生活中容易区分的类别。ImageNet-S_{300} 由 ImageNet-S_{50} 和250 个随机选取的类别组成。ImageNet-S_{50} 和 ImageNet-S_{300} 中图像的数量见表1。即使是 ImageNet-S_{50} 子集也比大多数语义分割数据集拥有更多的图像。

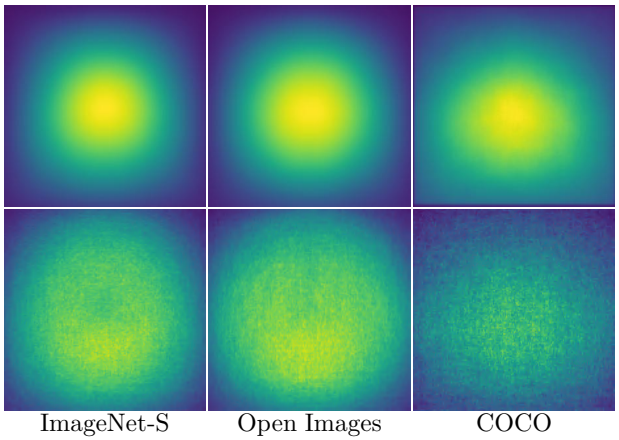


图 6. 数据集之间的物体位置分布比较：(top) **分割掩码的位置分布**，(down) **掩码边界的位置分布**。

3.2 评测

3.2.1 评测方案

因为在训练过程中缺少人为标注类别的监督，LUSS 模型不能像有监督学习得到的模型一样直接测试性能。因此，我们为 LUSS 提出了三个评测方案，包括完全无监督评测，半监督评测以及距离匹配评测。

完全无监督评测方案。完全无监督评测方案在训练期间不需要人为标注标签，只需要验证/测试集进行评测。与有监督任务不同，LUSS 任务中的类别是由模型生成的，在评测期间需要与 GT 类别匹配。我们提出了一个默认的图像级匹配方案，而更有效的匹配方案应该可以进一步提高 LUSS 评测性能。假设匹配集（通常为验证集）具有 N 张图像和 C 个类别。因为数据集有 C 个主要类别，因此类别的数量隐式包含在训练数据集中。我们假设无监督模型在训练过程中应该学会从数据集中生成超过 C 个类别。默认的图像级别匹配方案仅将 C 个生成的类别与 C 个真实类别相匹配。给定图像集 $\mathrm{D}=\{\mathrm{D}_k, k\in[1,N]\}$ 和 GT 标签 $\mathrm{G}=\{\mathrm{G}_k, k\in[1,N]\}$ 和预测的类别 $\mathrm{P}=\{\mathrm{P}_k, k\in[1,N]\}$，$\mathrm{G}_k$ 和 P_k 分别是图像 D_k 的 GT 和预测的类别集合。我们计算生成的类别和 GT 类别的匹配矩阵 $\mathrm{S}\in\mathbb{R}^{C\times C}$ 如下，其中 S_{ij} 表示在第 i 个生成的类别和第 j 个 GT 类别间的匹配度，当两个类别更可能是同一类别时其值更大：

$$\mathrm{S}_{ij}=\sum_{k=1}^{N}\mathbb{I}\{(i,j)\in \mathrm{P}_k\times \mathrm{G}_k\}, \tag{1}$$

$\mathrm{P}_k\times\mathrm{G}_k$ 是 P_k 和 G_k 的笛卡尔积，并且 (i,j) 属于 $\mathrm{P}_k\times\mathrm{G}_k$ 时 $\mathbb{I}$ 等价于1。利用匹配矩阵 $\mathrm{S}\in\mathbb{R}^{C\times C}$，我们在生成的类别和 GT 类别中使用匈牙利算法最大化 $\sum_{i=1}^{C}\mathrm{S}_{i,\mathrm{f}(i)}$ 找到了双射 $\mathrm{f}: i\mapsto j$。我们观察到有一些匹配错误的情况，并且一些生成的类别不在 GT 类别中，这表明了我们的基线匹配方法的局限性。我们希望未来的工作能够提出更有效的匹配方法来解决这个问题。

半监督评测方案。因为我们对大约 1% 的训练图像进行像素级标签标注，所以可以进行半监督微调以评测 LUSS 模型。半监督评测方案需要使用人工标注的训练数据对训练好的 LUSS 模型进行微调。因此，该方案不需要匹配生成的类别和 GT 类别。此外，该方案适用于现实世界中大量数据中只有部分图像被人工标注的现实应用场景。

距离匹配评测方案。在距离匹配评测方案中，我们直接利用像素级标注的训练图像获得 GT 类别的特征向量，并将其与验证/测试集中的特征向量进行匹配，以分配类别标签。具体来说，我们得到了训练集中每个类别的所有像素（包括“其他”类别），的平均特征向量和相应的类别标签。然后我们使用 k-NN 分类器 [84] 推理验证/测试集上的分割掩码。对于验证/测试集中的每个像素的特征向量，我们会在训练集中找到前 k 个相似的特征向量和相应的类别标签。每个像素的类别标签由这 k 个特征向量对应的类别投票决定。

3.2.2 评价指标

我们使用平均 IoU(mIoU), 边界 mIoU (b-mIoU), 图像级准确率 (Img-Acc), 和 F-measure (F_β) 作为 LUSS 任务的评价指标。在评测中，所有图像都使用原始图像分辨率进行评测。mIoU 和 b-mIoU 是综合评测指标，而 Img-Acc 和 F_β 分别从类别和形状方面评测模型性能。

mIOU. 类似于有监督语义分割任务 [8], [10], 我们使用 mIoU 来评测分割掩码的质量。除了主要类别外，“其他”类别也被用于计算 mIoU。

b-mIoU 与上述的评测所有物体区域的掩码的 mIoU 不同，b-mIoU [116] 重点关注边界区域。我们使用 b-mIoU 来评测边界区域的语义分割质量。根据在第3.1.2节中的分割一致性分析, 我们使用 $d=3\%$ 的 b-mIoU [116].

图像级准确率。 Img-Acc 可以评测模型的类别表征能力。由于许多图像包含多个标签，我们依照 [32] 将面积最大的预测类别是否属于该图片的 GT 类别集作为分类正确与否的评价标准。

F-measure。除了与类别相关的表征外，我们使用忽语义类别的 F_β 来评测形状质量 [117]。我们将主要类别视为前景类别，将“其他”类别视为背景类别。

4 一种 LUSS 方法

4.1 概述

我们总结了 LUSS 任务面临的主要挑战：1) 模型应该在无需图像级标签监督的情况下学习与类别相关的表征。2) 提取语义分割掩码需要模型学习形状表征。3) 形状和类别表征应在尽可能减少冲突的情况下共存。4) 利用学习到的表征，模型

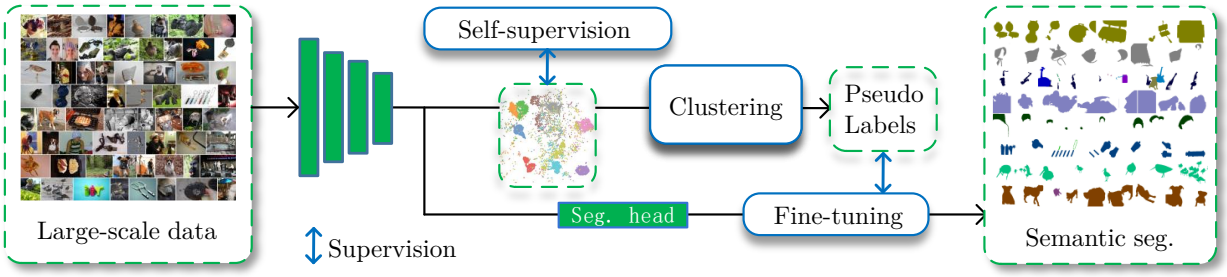


图 7. 本文提出的实现 LUSS 任务的其中一种流程。

应该高效地为图像中的每个像素分配自学习到的标签。5) 大规模的训练数据有助于以无监督学习的方式学习丰富的表征但不可避免地会消耗大量的训练成本，这就要求提高训练效率。

考虑到上述挑战，我们提出了一种新的 LUSS 方法，名为 PASS，(见图7)，包括四个步骤。1) 一个随机初始化的模型通过**自监督**的代理任务来学习形状和类别表征。经过表征学习，我们得到了所有训练图像的特征集。2) 然后，我们应用基于像素注意力的**聚类**方案来获得伪类别，并将生成的伪类别分配给每个图像像素。3) 我们用生成的伪标签**微调**预训练模型，以提高分割质量。4) 在**推理**时，LUSS 模型与有监督模型相同，即将生成的标签分配给图像的每个像素。注意，我们提出的流程不是 LUSS 任务唯一的选择，我们也鼓励未来工作使用其他的 LUSS 流程。下面我们详细介绍每个步骤。为了便于阅读，一些频繁使用的符号见表4。

4.2 无监督表征学习

对于我们的 LUSS 方法的第一步，一个随机初始化的模型，例如 ResNet，通过自我监督的代理任务来学习语义表征。LUSS 任务需要类别相关表征来区分不同类别的场景，并需要形状相关表征来构建物体的形状。之前的工作已经做了很多努力来学习图像级类别相关表征或像素级表征 [35], [94], [95]。然而，图像级方法通常忽略形状相关的特征。像素级方法更多关注有监督下游任务的迁移学习性能。通过 [118] 的发现，大多数下游任务的性能依赖于网络浅层的低级特征。因此，在下游任务中表现良好的像素级方法可能无法学习到有类别和形状信息的高级语义特征。

为了获得强大的表征来支持 LUSS 任务，我们提出了两种自监督学习策略来增强类别和形状表征，包括 1) 一种非对比像素到像素表征对齐策略，用于增强像素级形状相关表征，而不会损害实例级类别表征。2) 一种由深到浅的监督策略，以提高网络中间层特征的表征质量。

非对比像素到像素表征对齐。像素级形状相关表征旨在增强像素级的特征区分能力，即同一类别或来自同一图像的不同视图的相同位置的像素应具有一致的表征，反之亦然。我们观察到，大多数现有的像素级表征方法在 LUSS 任务上的性能比图像级表征方法差。我们认为现有的像素级方法过于关注像素级的区别，从而导致同一物体实例中像素间的语义差异。为了避免像素级表征对实例级类别表征的副作用，我们提出了一种非对比像素到像素表征对齐策略，该策略将来自同一图像的不同视图的相同位置的特征对齐，但不刻意增大不同位置的表征差异。

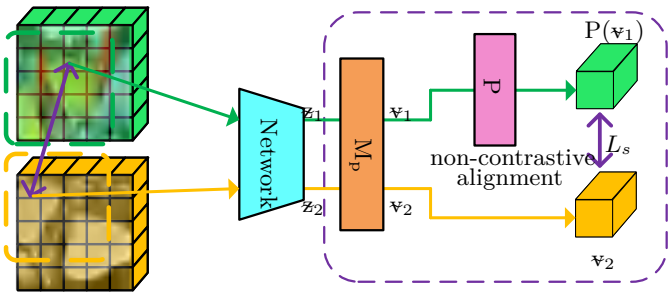


图 8. 非对比像素到像素表征对齐策略的图示。$\mathrm{M_p}$ 是确保像素级表征减小对类别表征干扰的映射层。P 是非对称损失的像素级预测器。

表 4. 常用符号的定义。

符号	维度/类型	含义
$\mathbf{z}$	$L \times H \times W$	一个图像的输出特征
$\mathbf{z}_k$	$L \times H \times W$	第 k 个图像的输出特征
$\mathbf{q}_k$	$(C+1) \times H \times W$	第 k 个图像的像素级伪标签
$\mathbf{y}_k$	$(C+1) \times H \times W$	第 k 个图像的像素级 GT 标签
C	scalar	主要类别数
L	scalar	输出特征的维度
H	scalar	输出特征的高度
W	scalar	输出特征的宽度
N	scalar	图像数
$\mathbb{P}$	operation	空间维度上的全局平均池化

如图8所示，给定从同一个图像的两个视图预测的特征对，我们在重叠像素提取特征图 $(\mathbf{z}_1, \mathbf{z}_2)$ 并且通过映射 $\mathbf{v} = \mathrm{M_p}(\mathbf{z})$ 获得像素级表征向量对 $(\mathbf{v}_1, \mathbf{v}_2)$，其中 $\mathrm{M_p}$ 是包含两个 1×1 卷积和激活层的像素级多层感知机 (MLP)。我们在第5.3.1节中展示，映射 $\mathrm{M_p}(\mathbf{z})$ 减少了像素级表征对类别表征的干扰。我们利用像素对像素对齐策略，使用非对称损失将两个视图重叠像素的特征向量对齐：

$$L_{I2I} = L_s(\mathrm{P}(\mathbf{v}_1), \mathscr{G}(\mathbf{v}_2)) + L_s(\mathscr{G}(\mathbf{v}_1), \mathrm{P}(\mathbf{v}_2)), \qquad (2)$$

其中映射 P 是像素级 MLP 预测器，$\mathscr{G}$ 是为了避免预测器崩

溃的停止梯度操作 [70], L_s 是余弦相似性损失函数。本文提出的非对比像素对像素对齐策略在不同视图之间形成稳健的像素级表征的同时保持了类别表征能力。

由深到浅的监督。网络浅层的低级、中级特征的质量，已被证明对视觉任务至关重要 [113], [119]。Islam et al. [119] 揭示了网络浅层中具有丰富的低/中级语义的表征，从而能够快速适应新任务。类似地, Kotar et al. [113] 展示了使用基于对比学习的方法能够有效学习高质量的低级特征。现有的大多数工作都是通过网络高层的间接梯度反向传播来优化中级表征 [25], [26], [33], [72]。我们观察到，由于低/中级特征缺乏语义信息，直接使用它们进行表征学习会导致次优性能。因此，我们提出了一种由深到浅的监督学习策略，以通过高质量高级特征监督的方式来增强低/中级特征的表征质量。

如图9, 给定从一幅图像经过数据增强得到的两个视图，我们从网络的 s 阶段得到特征对为 $(\mathrm{z}_1^{(s)}, \mathrm{z}_2^{(s)})$。为了简单起见，本文主要研究图像级由深到浅监督的影响。给定一个具有四个阶段的网络，用于由深到浅监督的图像级特征向量如下：

$$\mathrm{u}_i^{(s)} = \begin{cases} \mathrm{M}_I^s(\mathbb{P}(\mathrm{z}_i^{(s)})) & s = 4; \\ \mathrm{M}_I^s(\mathbb{P}(\mathrm{M}_K^s(\mathrm{z}_i^{(s)}))) & s < 4, \end{cases} \tag{3}$$

其中 $\mathbb{P}$ 是空间维度的全局平均池化操作，M_I^s and M_K^s 分别是阶段 s 的图像级/像素级 MLP 层。我们观察到直接全局池化中层特征会导致表征崩溃，因此添加 M_K^s 来避免此问题。在由深到浅的监督策略中，一个视图最后阶段的特征向量用于监督另一个视图的所有阶段的特征向量：

$$L_{D2S} = \frac{1}{|S|}\sum_{j}^{j\in S} L_I(\mathrm{u}_1^{(4)}, \mathrm{u}_2^{(j)}) + \frac{1}{|S|}\sum_{j}^{j\in S} L_I(\mathrm{u}_2^{(4)}, \mathrm{u}_1^{(j)}), \tag{4}$$

其中 S 是用于由深到浅监督的阶段的集合，L_I 是图像级损失。L_I 可以定义为多种形式，在本文中我们使用聚类损失 [33] 作为 L_I。

表征学习的训练损失。我们提出的像素到像素对齐和由深到浅监督可以与现有方法配合使用，以提高表征质量。无监督表征学习步骤的损失函数如下：

$$L_{sum} = L_{I2I} + L_{D2S} + L_e, \tag{5}$$

其中 L_e 是例如 SwAV [33] 和 PixelPro [35] 等现有方法的损失函数。

4.3 使用像素注意力生成像素标签

在表征学习之后，我们获得了所有训练图像的特征集合 $\mathrm{Z} = \{\mathrm{z}_k \in \mathbb{R}^{L\times H\times W}, k \in [1, N]\}$，其中 N 是图像的数量，L, H 和 W 是输出特征图的维度，高度和宽度。我们对 Z 进行聚类，以获得 C 个生成的类别，并将生成的类别分配给每个像素。标签生成的一种简单方法是对训练集中所有像素的特征向量进行聚类，LUSS 中的大规模数据导致聚类成本太高，例

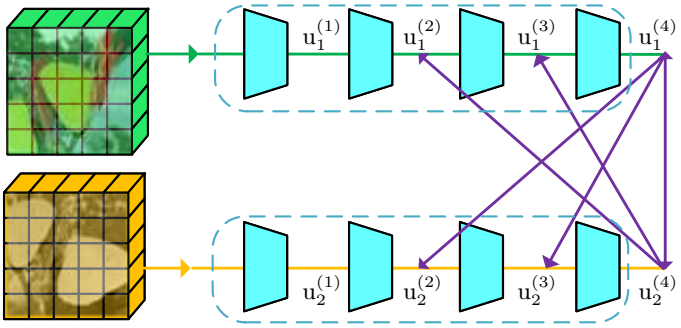


图 9. 由深到浅监督的图示。紫色线表征使用损失函数 L_I 作为监督。为了简单起见，省略了 M_I^s 和 M_K^s。

如聚类 ImageNet-S 训练集的 7×7 分辨率的像素级特征需要大约114 小时。另一种方法是使用在空间维度上聚类的图像级特征来节省聚类成本。然而，全局池化后的特征图中包含了许多不相关的特征信息，会影响聚类质量。

我们观察到，模型学习到的特征往往更关注语义更丰富的区域，例如具有更多有用语义信息的像素更有助于无监督表征学习模型的收敛。基于这一观察，我们提出了一种像素注意力方案，以突出有意义的语义区域，便于使用图像级特征生成像素级标签。具体来说，我们在模型的输出端添加一个像素注意力模块，并使用表征学习损失对其进行微调，以过滤掉语义信息较少的区域。具有像素注意力的过滤功能可以减少混合图像级特征向量中的噪声，从而提高聚类质量。此外，像素注意力将语义丰富的区域与语义较少的区域分开，从而在像素级标签生成过程中生成更精确的物体形状。我们在微调和标签生成步骤中给出了像素注意力的实现细节。

微调像素注意力。给定模型预测的一幅图像的特征 z，表征学习方法 [25], [33], 使用池化特征向量 $\mathrm{M}_I(\mathbb{P}(\mathrm{z}))$ 计算损失，其中 M_I 是图像级 MLP 层。池化操作平均地使用所有像素的特征，因为并非所有像素都表征有意义的语义，该操作不可避免地将噪声引入到图像级特征向量。本文的像素注意力被定义为：

$$\mathrm{c}(\mathrm{z}) = \sigma(\mathrm{M}_A(\|\mathrm{z}\|) + \theta), \tag{6}$$

其中 M_A 是像素级 MLP 层，$\theta \in \mathbb{R}^L$ 是初始化为 0 的可学习参数，σ 是限制输出注意值范围的 sigmoid 函数，$\|\mathrm{z}\|$ 是应用于特征 z 的通道维度的 L2 正则化操作。z 的每一个通道都有对应的像素注意力图。我们将像素注意力乘到特征图 z 上并且获得像素注意力增强的图像级特征向量 $\hat{\mathrm{v}} = \mathrm{M}_I(\mathbb{P}(\mathrm{c}(\mathrm{z})\cdot\|\mathrm{z}\|))$。在微调过程中，我们将回传到网络的梯度断开，只利用 $\hat{\mathrm{v}}$ 计算的表征损失优化像素注意力模块。我们发现使用聚类损失 [33] 微调的像素注意力模块能获得与形状相关的像素注意力图（见图10）。

基于像素注意力的标签生成。基于像素注意力 c(z), 我们得到像素注意力增强的图像级别特征 $\hat{\mathrm{Z}} = \{\hat{\mathrm{Z}}_k \in \mathbb{R}^L, k \in [1, N]\}$,

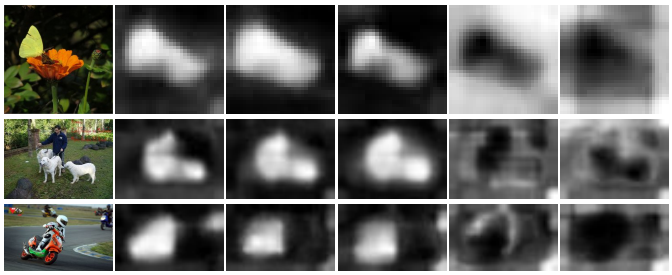


图 10. 不同通道像素注意力图的可视化。大多数像素注意力图突出显示语义区域，而一些通道突出显示背景区域。

其中 $\widehat{\mathrm{Z}}_k = \mathbb{P}(\mathrm{c}(\mathrm{z})_k \cdot \|\mathrm{z}_k\|)$。我们在 $\widehat{\mathrm{Z}}$ 上构造 k-means 聚类来生成 C 个类别的聚类中心 $K \in \mathbb{R}^{L\times C}$。根据生成的类别，我们需要给图像分配像素级伪标签 $\mathrm{Q} = \{\mathrm{q}_k \in \mathbb{R}^{C+1\times H\times W}, k \in [1, N]\}$。我们在图10中展示了微调后的像素注意力可以突出图像中的语义区域。因此，我们基于像素注意力提取语义信息丰富的区域：

$$\mathrm{d}(\mathrm{z}) = \begin{cases} 0 & \frac{1}{L}\sum_i^{0\le i<L} \mathrm{c}(\mathrm{z})_i < \tau; \\ 1 & \frac{1}{L}\sum_i^{0\le i<L} \mathrm{c}(\mathrm{z})_i \ge \tau, \end{cases} \tag{7}$$

其中 τ 是用于区分主要类别和“其他”类别的预定义的阈值，低于 τ 的区域被视为“其他”类别。对于主要类别区域中的每个像素，我们将聚类中心 K 中与该像素的特征向量距离最小的类别作为其类别。

4.4 微调和推理

在微调步骤中，我们加载无监督表征学习得到的预训练权重并添加具有 $L \times (C + 1)$ 个通道的 conv 1×1 层作为分割头。使用伪标签 Q 用交叉熵损失来监督分割头的输出特征 $\mathrm{Y} = \{\mathrm{y}_k \in \mathbb{R}^{(C+1)\times H\times W}, k \in [1, N]\}$ 从而微调模型。LUSS 模型的推理过程与一个完全有监督的语义分割模型的推理过程一致。对于每个在 y_k 中的像素特征向量 $\mathrm{w} \in \mathbb{R}^{C+1}$，通过如下式子获得分割类别标签：

$$\mathrm{w} = \arg\max_{i\in[1,C+1]}(\mathrm{w}_i). \tag{8}$$

5 实验和分析

5.1 实现细节

表征学习步骤的训练细节。我们在 ImageNet-S_{50} 数据集上使用 ResNet-18 网络，在 ImageNet-S_{300} 和 ImageNet-S 数据集上使用 ResNet-50 网络。为了公平比较，所有网络使用256 的批大小，在 ImageNet-S_{50} 上训练 200 个迭代轮次，在 ImageNet-S_{300}/ImageNet-S 上训练 100 个迭代轮次。

我们分别基于图像级方法 SwAV [33] 和像素级方法 PixelPro [35] 实现了本文提出的表征学习方法。依照 SwAV [33]，LARS 优化器用于更新网络，权重衰减为 1e-6，动量为 0.9。初始学习速率为 0.6，并使用余弦学习率调整策略逐渐衰减到 6e-6。对于 ImageNet-S_{300} 和 ImageNet-S 数据集，为与其他方法进行公平比较，我们只使用大小为 224×224 的两个裁剪视图进行训练，而没有使用 multi-crop 训练策略 [33]。与 SwAV 相同，从第15 个迭代轮次开始使用一个长度为3,840 的队列，并且在5,005 次迭代前将聚类中心冻结。当在 ImageNet-S_{50} 上进行训练时，为加速收敛，队列长度设置为 2048，并且聚类中心在1,001 次迭代之前被冻结。在 ImageNet-S_{50} 数据集上训练，我们使用 multi-crop 训练策略，其中包含六个大小为 96×96 的裁剪视图，两个裁剪大小为 224×224 的视图。当和 PixelPro [35] 结合时，训练模式与官方设置一致。我们使用 LARS 优化器训练网络，初始学习率为 1.0。经过五个预热迭代轮次后，学习率随着余弦学习率调整策略逐渐下降到 1e-6。

微调步骤的训练细节。为了生成像素级标签，我们首先对像素注意力模块微调20 个迭代轮次，此时固定在无监督表征学习阶段训练好的模型参数。我们默认使用聚类损失 [33] 来微调像素注意力模块。训练策略与表征学习步骤中的策略相同。

在微调步骤中，表征学习损失被移除，使用交叉熵损失来监督分割头。我们加载在表征学习步骤上预先训练好的模型权重，并对其进行20 个迭代轮次的微调。我们使用一个权重衰减为 1e-6、批大小为 256、动量为 0.9 的 LARS 优化器来训练网络。初始学习速率为 0.6，随着余弦学习率调整策略逐渐衰减到 6e-6。

5.2 与 USS 方法比较

在本节中，我们使用完全无监督评测方案在 ImageNet-S 数据集上评测本文提出的 LUSS 方法的性能。表5表明我们的方法在大规模数据上取得了合理的性能。图11中的可视化说明大规模数据的无监督语义分割是可行的。

与无监督语义分割方法的比较。现存的无监督语义分割 (USS) 方法被设计用于相对较小规模的数据，因此由于训练时间限制不能被直接用于 ImageNet-S 数据集。因此，我们在 ImageNet-S_{50} 子集上将我们的 LUSS 方法和现有的 USS 方法进行比较，见表5。为了公平比较，在 ImageNet-S_{50} 数据集上训练的方法都使用了 ResNet-18 网络。严格来说，这种比较并不公平，因为一些现有的 USS 方法没有在完全无监督的条件中进行训练。例如, MDC [88] 和 PiCIE [46] 使用有监督 $\mathrm{ImageNet}_{1k}$ 预训练权重初始化模型。这两个方法在使用 MoCo [25] 预训练权重时会有大幅度性能下降，这说明有监督预训练对这些方法必不可少。MaskContrast [31] 使用 MoCo 预训练权重初始化模型并使用额外的显著图作为监督进行训练。如果从头开始训练此模型，将导致很大的性能损失。相反，我们的 LUSS 方法是从头开始训练的，没有使用直接或间接的人为监督信息。我们的方法包括新的表征学习策略、标签生成方法和微调方案。为了验证我们方法的通用性，我们基于两种表征学习方法来实现我们的方法，i.e., SwAV [33] 和 PixelPro

表 5. 在 ImageNet-S 数据集上使用完全无监督评价方案下我们提出的 LUSS 方法和现有的 USS 方法的比较。不同物体尺寸下的测试 mIoU 也在表中提供。† 代表从头开始训练模型 200 个迭代轮次。本文方法的 $_{s/p}$ 分别表示在 (5)中使用 SwAV [33] 和 PixelPro [35] 作为 L_e。S 代表该方法使用显著图。I 代表使用有监督 ImageNet$_{1k}$ 预训练权重作为模型初始化。默认情况下，“其他”类别用于计算 mIoU 和 b-mIoU。我们在补充材料中也给出了不考虑“其他”类别的性能对比。

LUSS	Pior	mIoU		b-mIoU		Img-Acc		F_β		mIoU				b-mIoU			
		val	test	val	test	val	test	val	test	S.	M.S.	M.L.	L.	S.	M.S.	M.L.	L.
ImageNet-S$_{50}$																	
MDC [46], [88]	-	4.0	3.6	1.4	1.2	14.9	13.4	31.6	31.3	0.4	2.6	3.8	4.9	0.2	1.1	1.4	1.5
MDC [46], [88]	I	14.6	14.3	3.1	3.1	44.8	40.8	33.2	32.6	2.6	10.9	14.6	19.1	0.9	2.2	3.2	4.7
PiCIE [46]	-	5.0	4.5	1.8	1.6	15.8	14.0	14.6	32.2	0.2	3.1	5.0	5.3	0.2	1.2	1.7	1.9
PiCIE [46]	I	17.8	17.6	3.7	4.0	45.0	44.0	32.1	31.6	4.4	13.1	20.1	23.1	1.0	2.7	4.4	5.8
MaskCon [31]	S	24.6	24.2	15.6	15.1	47.9	47.6	65.7	66.2	12.2	25.6	24.7	20.4	10.1	17.0	14.5	10.6
MaskCon [31]†	S	13.9	10.5	8.5	10.5	30.2	22.4	62.6	62.3	2.5	2.1	1.7	1.7	2.4	6.3	6.5	5.7
PASS$_s$	-	29.2	29.3	7.6	7.4	66.2	65.5	49.0	49.0	6.6	25.0	33.2	32.6	3.3	6.2	8.1	9.5
PASS$_p$	-	32.4	32.0	7.2	7.2	62.9	64.1	48.7	47.9	9.7	26.2	36.5	40.5	5.1	5.8	7.8	10.4
PASS$_p$+ RC [117]	S	42.6	42.1	17.5	17.7	58.8	61.8	62.1	61.3	17.0	38.6	45.5	43.7	11.2	17.2	19.0	17.1
PASS$_p$+ Sal	S	43.3	42.3	20.4	20.2	64.6	65.2	70.0	69.9	19.0	41.7	45.1	38.3	14.7	22.6	20.6	15.3
ImageNet-S$_{300}$																	
PASS$_p$	-	16.6	16.0	4.4	4.2	34.7	32.8	34.4	34.3	2.8	12.0	16.4	21.7	1.4	3.2	3.9	6.4
PASS$_s$	-	18.0	18.1	5.2	5.2	43.9	42.6	47.6	47.5	4.2	13.6	19.5	23.5	2.1	4.2	5.5	7.1
ImageNet-S																	
PASS$_p$	-	7.3	6.6	2.4	2.1	19.9	18.0	34.8	34.6	1.3	4.6	7.1	8.4	0.6	1.5	2.1	2.8
PASS$_s$	-	11.5	11.0	3.8	3.5	24.0	22.3	37.1	36.9	2.4	8.3	11.9	13.4	1.3	3.0	3.8	4.3

[35]。我们的方法在 mIoU 指标上相比现有的 USS 方法有显著提升。得益于额外显著图监督，MaskContrast 比我们的方法有更高的 F_β。当我们的方法也使用相同的显著图，在 F_β 指标明显优于 MaskContrast 并且达到了更高的 mIoU。注意在 [31] 中的显著图不是严格的无监督版本，因为其使用了有监督的 ImageNet 预训练权重。我们还实现了其他 USS 方法，例如 IIC [90]。然而，由于这些方法是为仅有几个类别的简单语义分割而设计的，因此它们无法在 ImageNet-S$_{50}$ 数据集收敛。

不同尺寸物体的性能。如第3.1.2节中介绍，ImageNet-S 数据集根据物体大小被分为不同组。我们评测了不同物体尺寸下的 test mIoU，见表5。

在 mIoU 和 b-mIoU 指标下，小物体的性能比大物体差，这表明小物体需要一个具有更精确像素级表征和分割能力的模型。注意，b-mIoU 中不同物体大小的性能差异比 mIoU 小，因为 b-mIoU 对物体大小变化更为鲁棒。

不同数据规模之间的差异。如表5，我们在 ImageNet-S$_{50}$，ImageNet-S$_{300}$，和 ImageNet-S 数据集上训练我们的模型。随着数据规模的增长，模型性能下降，展现了大规模数据对无监督语义带来的挑战。我们观察到，在 ImageNet-S$_{50}$ 数据集上，基于 PixelPro 的方法优于基于 SwAV 的方法，但基于 SwAV 的方法在 ImageNet-S300 和 ImageNet-S 数据集上取得了更好的性能。我们认为不同的数据量偏好不同的表征学习策略。

图 11. 无监督语义分割结果的可视化. 最后三行在标签生成过程中使用显著的先验信息进行训练，显示出更好的形状质量。

为了评测大数据集和小数据集之间的性能差异，我们在大数据集上训练模型，在小数据集上评测模型，见表6。在大数据集上训练的模型性能不如在小数据集上训练的模型。在 ImageNet-S$_{50}$ 数据集上测试，在 ImageNet-S$_{50}$ 数据集上训练的模型在所有指标都达到了最好的表现，而在 ImageNet-S$_{919}$

表 6. 使用 PASS$_s$ 方法和完全无监督的方法，在 ImageNet-S 完整数据集上训练模型，在 ImageNet-S 子集上评测模型。

Training set	mIoU		Img-Acc		F_β	
	val	test	val	test	val	test
Testing on ImageNet-S$_{50}$						
ImageNet-S$_{50}$	29.2	29.3	66.2	65.5	49.0	49.0
ImageNet-S$_{300}$	27.8	27.4	65.2	63.3	38.6	36.0
ImageNet-S	24.1	23.0	61.3	57.8	31.3	28.9
Testing on ImageNet-S$_{300}$						
ImageNet-S$_{300}$	18.0	18.1	43.9	42.6	47.6	47.5
ImageNet-S	16.4	16.6	39.3	37.2	36.8	36.0

表 7. 本文提出的 P2P 对齐和 D2S 监督表征学习策略的消融实验。所有模型训练100 迭代轮次。D2S3 和 D2S32 分别表示监督网络的第 3 和第 2-3 阶段。

(a) 使用距离匹配评测方案的 LUSS 消融实验。

ImageNet-S$_{300}$	mIoU		Img-Acc		F_β	
	val	test	val	test	val	test
SwAV [33]	22.4	22.6	57.4	57.5	63.5	63.7
+P2P	24.8	24.8	58.4	58.5	64.5	64.8
+P2P-D2S3	25.1	25.2	57.3	57.5	65.0	65.2
+P2P-D2S32	24.8	24.9	56.8	56.6	65.7	66.0
PixelPro [35]	15.5	15.8	44.0	44.3	62.4	62.6
+Clustering Loss	20.8	21.3	52.0	52.1	61.5	62.1
+P2P	21.3	22.0	52.2	52.8	61.5	62.1
+P2P-D2S3	22.2	22.8	53.2	53.1	62.2	62.9
+P2P-D2S32	23.0	23.4	53.3	54.3	62.4	63.1

(b) 下游任务迁移学习的消融实验。

ImageNet-S$_{300}$	COCO SEG			COCO DET			VOC SEG
	AP	AP50	AP75	AP	AP50	AP75	mIoU
SwAV [33]	32.4	52.1	34.6	35.5	54.9	38.6	68.9
+P2P	32.8	52.5	34.9	36.0	55.4	39.1	70.4
+P2P-D2S3	33.5	53.4	35.8	36.7	56.4	39.4	70.8
+P2P-D2S32	33.8	53.7	36.2	37.2	56.6	40.6	70.8
PixelPro [35]	34.7	54.8	37.2	38.2	57.5	41.7	72.8
+Clustering Loss	34.9	55.2	37.3	38.4	58.1	41.9	73.3
+P2P	35.3	55.9	37.9	38.9	58.6	42.4	72.3
+P2P-D2S3	35.3	55.9	37.6	38.8	58.6	42.3	73.9
+P2P-D2S32	35.7	56.6	38.3	39.4	59.1	43.1	75.1

数据集上训练的模型有最差的性能。在 ImageNet-S$_{300}$ 数据集上进行评测时，也观察到类似的趋势。这些结果表明，在大数据集上训练无监督模型比在小数据集上更难，这表明了大规模数据的巨大挑战。然而，这些性能差距相对较小，可以通过未来更强的方法来弥补。

5.3 消融实验

5.3.1 表征学习

在本节中，我们在 LUSS 任务上对我们提出的和一些现有的无监督表征学习方法进行了基准测试。除非另有说明，否则我们使用 ImageNet-S$_{300}$ 数据集进行实验以节省计算成本。为了避免 LUSS 中微调步骤的影响，我们用第3.2.1节中介绍的距离匹配评测方案来评测 LUSS 方法。

本文提出的表征学习方法的消融实验。我们在 SwAV [33] 和 PixelPro [35] 上实现了提出的非对比像素到像素（P2P）对齐和由深到浅（D2S）监督。表7(a) 展示了 PixelPro 由于其缺少 LUSS 任务所需要的类别相关表征能力，效果比 SwAV 差。因此，我们在 PixelPro 中添加了聚类损失 [33] 来针对 LUSS 任务构建一个合理的基准。如表7(a)，我们的方法在 ImageNet-S$_{300}$ 数据集上分别相较于 SwAV 和 PixelPro 的 test mIoU 提升了 2.6% 和 7.6%。具体而言，与图像级方法 SwAV 相比，P2P 对齐在 test mIoU 中的增益为 2.2%，在 test mOU 中，它还将使用聚类损失增强后的 PixelPro 提高了 0.5%。与基于 SwAV 和 PixelPro 的基准相比，D2S 监督分别带来 0.4% 和 1.4% 的进一步提升。总之，P2P 对齐有效地增强了图像级方法的像素级表征，D2S 监督丰富了像素级方法的实例级类别表征。P2P 对齐和 D2S 监督分别改进了像素级和图像级表征方法，展示了本文所提策略的鲁棒性。如表9，我们提出的表征学习策略在 ImageNet-S 数据集上也比基准要效果更好。

非对比像素对像素对齐。我们利用非对比 P2P 对齐来增强像素级表征，而不会损害实例级类别表征。我们还比较了不同的像素级对齐策略，包括聚类、对比和非对比类型。对于聚类和对比度损失，我们将两个视图相同位置的像素设置为正样本，其他像素设置为负样本。如表8(a)，与基准相比，这两种像素级对齐策略都具有更高的 F_β，显示出形状表征质量的提升。然而，由于同一物体中像素之间的语义差异，聚类和对比度损失的 mIoU 和 Img-Acc 性能较差。相比之下，由于保持了属于同一语义实例的像素的表征一致性，本文提出的非对比 P2P 对齐在 mIoU 和 Img-Acc 中的性能优于基准方法。我们也在表8(b)中分析了映射 $\mathrm{M_p}(\cdot)$ 的有效性。有映射 $\mathrm{M_p}(\cdot)$ 的 P2P 对齐可以实现更好的 Img-Acc 性能，因为 $\mathrm{M_p}(\cdot)$ 减小了像素级表征对类别相关表征的干扰。

由深到浅的监督。 D2S 监督利用最后阶段的高质量特征来监督早期特征。表8(c)比较使用相同或最后阶段的特征作为监督浅层。我们观察到，这两种设置都比基准有所提升，由深到浅的监督在 mIoU 和 Img-Acc 上优于同一阶段监督。默认情况下，我们使用从一个视图的深层特征来监督另一个视图的浅层特征。如表8(d)，我们研究了使用来自同一视图的特征对 D2S 监督的影响。交叉视图监督略优于相同视图监督。我们观察到，相同视图监督的训练损失低于交叉视图监督。我们

表 8. 使用距离匹配评测方案在 ImageNet-S_{300} 测试集对 P2P 对齐和 D2S 监督策略的消融实验。

(a) P2P 对齐的不同损失函数形式。

ImageNet-S_{300}	mIoU	Img-Acc	F_β
SwAV baseline	22.6	57.5	63.7
+Clustering P2P	21.2	51.8	66.4
+Contrastive P2P	18.0	46.4	64.6
+Non-contrastive P2P	24.8	58.5	64.8

(b) M 映射在 P2P 对齐的作用。

ImageNet-S_{300}	mIoU	Img-Acc	F_β
SwAV baseline	22.6	57.5	63.7
P2P without $M_p(\not{z})$	24.6	57.1	64.9
P2P with $M_p(\not{z})$	24.8	58.5	64.8

(c) D2S 监督中由深到浅层监督与同层监督的对比。

ImageNet-S_{300}	mIoU	Img-Acc	F_β
PixelPro+P2P (baseline)	22.0	52.8	62.1
+same-stage sup.	22.6	52.9	63.1
+deep-to-shallow sup.	23.4	54.3	63.1

(d) D2S 监督中使用同一个视图和不同视图的特征进行监督。

ImageNet-S_{300}	mIoU	Img-Acc	F_β
PixelPro+P2P (baseline)	22.0	52.8	62.1
+same-view sup.	23.1	53.9	63.2
+cross-view sup.	23.4	54.3	63.1

认为，相同视图监督会导致训练过拟合，影响测试性能。D2S 监督可以用于网络不同浅层阶段的多个特征。如表7(a)，我们研究了监督网络不同阶段特征对基于 SwAV 和 PixelPro 的方法的影响。我们发现不同的方法需要监督不同的阶段来获得最优结果，例如在 SwAV 中监督阶段 3-2 比监督阶段 3 效果差，但 PixelPro 能从更多的对浅层的监督中获益。我们通过消融使用决定 D2S 监督的阶段。

评测无监督学习方法。为了分析无监督学习方法在 LUSS 任务中的表征能力，我们对包括对比、非对比、聚类和像素级等有代表性的方法进行分类并进行基准测试。如表9，图像级方法在 mIoU，Img-Acc，F_β 上比像素级方法有更明显的优势。像素级方法过于关注像素级特征的差异，导致同一物体实例中像素之间的语义差异较大。相比之下，图像级方法提供了一致的实例级类别相关表征，因为这些方法的损失函数重点优化模型对图像之间的区分能力。然而，像素级表征对 LUSS 任务至关重要，因为我们提出的非对比 P2P 对齐方法相比于图像级方法 SwAV 有相当大的提升。我们发现聚类方法在 Img-Acc 上比对比和非对比方法都要高，但在形状相关

表 9. 本文的无表征学习增强方法与其他无监督表征学习方法在距离匹配评测方案下的性能比较。我们的 $_{s/p}$ 分别表示使用 SwAV [33] 和 PixelPro [35] 作为(5)中的 L_e。所有模型训练100 个迭代轮次。Supervised 表示使用图像级有监督预训练权重初始化模型。

LUSS	mIoU		Img-Acc		F_β	
	val	test	val	test	val	test
ImageNet-S_{300}						
Supervised	33.8	33.9	80.4	81.5	60.0	60.0
Contrastive						
SimCLR [26]	12.5	12.6	37.7	38.4	63.7	64.0
MoCov2 [25], [120]	12.4	12.4	40.3	40.3	64.1	64.4
AdCo [121]	21.1	21.5	55.1	54.8	64.9	65.5
Non-contrastive						
BYOL [34]	13.4	13.4	38.3	38.0	64.0	64.4
SimSiam [70]	20.1	20.3	56.9	57.5	65.5	66.0
Clustering						
PCL [93]	17.4	17.9	48.4	48.0	63.0	63.3
SwAV [33]	22.4	22.6	57.4	57.5	63.5	63.7
$PASS_s$	25.1	25.2	57.3	57.5	65.0	65.2
Pixel-level						
DenseCL [95]	13.9	13.8	36.4	36.8	63.7	63.7
PixelPro [35]	15.5	15.8	44.0	44.3	62.4	62.6
$PASS_p$	23.0	23.4	53.3	54.3	62.4	63.1
ImageNet-S						
Supervised	30.0	29.8	75.9	76.6	58.7	58.7
PixelPro [35]	7.7	7.5	26.9	26.5	61.8	61.8
$PASS_p$	9.8	9.8	29.4	29.6	61.1	61.3
SwAV [33]	15.1	15.1	43.5	43.3	64.2	64.3
$PASS_s$	15.6	15.6	43.1	42.9	64.3	64.6

的 F_β 上要低。与对比和非对比方法相比，聚类方法鼓励使用聚类中心进行更强的类别相关表征学习。但由于聚类方法中一幅图像的所有像素都接近类别质心，因此主类别与“其他”类别之间的表征差异会减弱。图像级有监督方法比聚类方法有更好的类别中心，但在 F_β 上更差。这些结果解释了为什么聚类方法有更差的 F_β。

类别在 LUSS 任务中扮演什么角色？为了回答这个问题，我们使用经过图像级有监督训练的模型作为基准。如表9，在 mIoU 指标上，有监督模型的性能优于无监督模型。此外，它在图像级分类精度方面大大优于无监督模型。相反，它在形状相关指标，例如 F_β，比大多数无监督方法要差。这些结果表明，类别特征确实有助于 LUSS 任务。然而，形状特征不能仅通过类别表征学习来学习。

使用完全无监督评测方案在 ImageNet-S_{50} 测试集上对像素标签生成和微调步骤进行消融。

表 10. 使用完全无监督评价方案在 ImageNet-S_{50} 测试集的像素级标签生成和微调步骤的消融实验。

(a) 不同伪标签生成方式的对比。$^{\tau}$ 代表使用图像级方法的推理策略。

ImageNet-S_{50}	mIoU	Img-Acc	F_β
Image-level	26.9	57.6	53.0
Pixel-level	12.7	37.4	32.9
Pixel-attention	29.3	65.5	49.0
Pixel-attention$^{\tau}$	29.2	61.7	52.3

(b) 不同伪标签生成方法的聚类时间 (秒)。

	ImageNet-S_{50}	ImageNet-S_{300}	ImageNet-S
Image-level	2.8×10^0	8.9×10^1	7.5×10^2
Pixel-level	3.2×10^2	4.6×10^4	4.1×10^5
Pixel-attention	2.8×10^0	8.9×10^1	7.5×10^2

(c) 对输出特征是否共享像素注意力图。

ImageNet-S_{50}	mIoU	Img-Acc	F_β
Shared	28.4	64.3	48.8
Unshared	29.3	65.5	49.0

(d) 微调步骤对 LUSS 方法的性能影响。

ImageNet-S_{50}	mIoU	Img-Acc	F_β
Before fine-tuning	26.0	63.8	44.7
After fine-tuning	29.3	65.5	49.0

5.3.2 标签生成和微调

我们使用第3.2.1节节中描述的完全无监督评测方案评测本文提出的基于像素注意力的标签生成和微调方案的有效性。除非另有说明，否则我们使用 ImageNet-S_{50} 子集进行消融实验。

像素标签生成的效果。我们将提出的基于像素注意力的像素标签生成方法与图像级和像素级标签生成方法进行了比较。我们首先简要介绍使用图像级和像素级标签生成方法的标签生成和微调过程。图像级别标签生成方法在池化图像级别特征向量上聚类出 C 个类别，并将图像级别标签分配给每个图像。在微调期间，使用图像级标签监督全连接（FC）层。为了获得像素级分割掩码，在推理阶段 FC 层被替换为 1×1 卷积层。由于缺少“其他”类别，我们采用 WSSS 方法广泛使用的基于类激活映射（CAM）的掩码生成方法来生成最终的分割掩码。附录中介绍了实施细节。在大规模 ImageNet-S 数据集上，像素级特征向量聚类的成本太高。相反，我们在 ImageNet-S_{50} 数据集上实现像素级方法来进行比较。我们使用像素级特征向量对 $C+1$ 类别进行聚类，并使用像素级标签对其进行微调。如表10(a)，本文提出的基于像素注意力的标签生成方法优于图像级和像素级方法，且具有相当大的优势。图像级别方法比我们的方法具有更好的 F_β。我们的方法在应用这种推理策略后，F_β 也得到了显著改进，而 mIoU 的变化可以忽略不计。

聚类时间比较。我们将基于像素注意力的标签生成的聚类时间与表10(b)中的其他两种标签生成方法进行了比较。我们的方法与图像级方法具有相同的聚类时间，因为它们都使用图像级特征向量。使用低分辨率 7×7 的输出特征图，像素级方法比我们的方法慢得多，因为训练集中有大量的像素。在完整的 ImageNet-S 数据集上进行聚类时，像素级方法的时间大约为114 小时，这对于实际使用来说是不可接受的。

输出特征共享/非共享像素注意力图。默认情况下，我们为每个输出特征通道生成一个的像素注意力图。我们还研究了对所有通道使用一个共享的像素注意力图的效果。在表10(c)中的结果说明对每个通道使用非共享像素注意力图可以获得更好的性能。我们在图10中可视化了不同通道的像素注意力图。大多数通道侧重于语义区域，而少数通道突出显示背景区域。此外，每个像素注意力图的焦点也不相同，这说明了非共享像素注意力的有效性。

微调的效果。我们基于像素注意力的标签生成方法可以直接生成像素级分割掩码。我们比较了微调步骤前后的性能，以验证我们的 LUSS 方法中微调步骤的效果。如表10(d)，微调方法将 test mIoU 提升了 3.3%，说明生成的像素级标签仍然有噪声，微调进一步提高了语义分割质量。

5.4 下游任务的迁移学习

在 LUSS 任务提出之前，自监督表征学习方法主要用作下游任务迁移学习的预训练 [25], [35]。LUSS 任务需要来自自监督表征学习的形状相关表征和类别相关表征。在本节中，我们研究 LUSS 任务学习的表征是否有利于像素级下游任务，例如语义分割、实例分割和目标检测。我们还比较了不同表征学习方法对 LUSS 和下游任务的影响。为了公平比较，除非另有说明，不同的表征学习方法均使用 ResNet-50 [108] 网络在 ImageNet-S_{300} 或 ImageNet-S 数据集上预训练100 个迭代轮次。

实例分割和目标检测。我们使用 MaskRCNN [27] 作为实例分割和目标检测的检测器。模型在 COCO17 [107] 训练集上训练，并且在验证集上评测。依照通用设置 [25], [27], [35]，我们加载在不同表征学习方法上预训练的 ResNet-50 的权重，并应用 1× 训练计划。如表11所示，我们基于 SwAV [33] 和 PixelPro [35] 验证了本文提出的非对比 P2P 对齐和 D2S 监督表征学习策略。我们首先比较在 ImageNet-S_{300} 数据集上预训练模型的性能。在实例分割中，我们的方法在 mAP 指标上相较 SwAV 和 PixelPro 分别提升 1.4% 和 1.0%。类似地，目标检测的 mAP 分别提升 1.7% 和 1.2%。这些结果证明，我

表 11. 在对 ImageNet-S_{300} 和 ImageNet-S 数据集预训练的无监督表征学习方法之间的迁移学习比较。所有模型训练100 个迭代轮次. 我们的 $_{s/p}$ 分别表征使用 SwAV [33] 和 PixelPro [35] 作为 (5)中的 L_e。Supervised 表示通过图像级有监督预训练权重对模型进行初始化。

Transfer learning	COCO SEG			COCO DET			VOC SEG
	AP	AP50	AP75	AP	AP50	AP75	mIoU
ImageNet-S_{300}							
Supervised	34.7	55.3	37.0	38.4	58.1	42.0	72.6
Contrastive							
SimCLR [26]	31.9	51.1	34.1	35.0	53.7	38.2	66.4
MoCov2 [25], [120]	33.7	53.6	36.1	37.1	56.3	40.3	67.8
AdCo [121]	34.3	54.3	36.7	37.9	57.2	41.5	70.0
Non-contrastive							
BYOL [34]	32.1	51.6	34.2	35.1	54.2	38.2	65.8
SimSiam [70]	33.7	53.3	36.2	36.9	56.0	40.3	61.1
Clustering							
PCL [93]	34.3	54.4	36.9	37.8	57.0	41.3	69.6
SwAV [33]	32.4	52.1	34.6	35.5	54.9	38.6	68.9
$PASS_s$	33.8	53.7	36.2	37.2	56.6	40.6	70.8
Pixel-level							
DenseCL [95]	33.7	53.4	36.2	37.0	56.2	40.4	67.7
PixelPro [35]	34.7	54.8	37.2	38.2	57.5	41.7	72.8
$PASS_p$	35.7	56.6	38.3	39.4	59.1	43.1	75.1
ImageNet-S							
Supervised	36.6	57.5	39.4	40.3	60.5	44.0	76.4
SwAV [33]	34.4	55.0	36.8	37.8	58.0	41.1	73.0
$PASS_s$	35.3	56.0	37.8	38.9	58.8	42.3	75.3
PixelPro [35]	35.9	56.6	38.6	39.5	59.2	43.1	73.9
$PASS_p$	36.5	57.4	39.1	40.2	60.3	44.1	76.1

们用于 LUSS 任务的表征学习方法在实例分割和目标检测任务的不同基准上具有稳定的性能增益。像素级方法 PixelPro 优于例如 SwAV、AdCo 和 SimSiam 等其他图像级方法，这证明像素级方法对这两个像素级下游任务具有更强的迁移能力。当使用完整的 ImageNet-S 数据集进行预训练时，我们的方法仍然优于基线，例如基于 PixelPro 的本文方法在实例分割和目标检测任务中分别获得 0.6% 和 0.7% 的 mAP 增益。

语义分割。我们还使用基于 ResNet-50 的 Deeplab V3+ [12] 网络将预训练好的模型权重迁移到 PASCAL VOC 数据集 [123] 上的语义分割任务。模型在 Pascal VOC SBD 训练集上训练并在验证集上评测。依照 [124] 的训练设置，我们用16 的批次大小迭代训练 20k 次。图像按 0.5 到 2.0 的比例缩放，然后裁剪为512 边长进行训练。在 ImageNet-S_{300} 数据集上进行预训练时，我们的方法在 mIoU 上比 SwAV 和 PixelPro 分别高了 1.9% 和 2.3% 。使用 ImageNet-S 预训练模型时性能在 mIoU 上分别提升了 2.3% 和 2.2%。像素级方法 PixelPro 与其他图像级方法相比具有明显优势，表明像素级表征对于语义分割至关重要。基于对比学习的方法虽然仍是图像级方法，但在语义分割方面优于聚类和非对比学习方法。

LUSS 与迁移学习的关系。我们在表9和表11中分别比较了表征学习方法在 LUSS 和下游任务上的性能。与图像级方法相比，SwAV 聚类方法由于分类精度高，在 LUSS 任务中具有更好的性能。在下游任务上，SwAV 不如许多在 LUSS 任务上性能较差的方法。例如，在下游实例分割任务中，对比学习方法 MoCov2 的 mAP 比 SwAV 提升 1.3%，但 LUSS 任务的 mIoU 有 10% 的差距。这一观察结果与 et al. [113] 的发现一致，对比方法可以更好地学习低层特征，从而有利于像素级下游任务。与图像级方法相比，像素级方法 PixelPro 明显优于图像级方法。但在 LUSS 任务中，它的性能比许多图像级方法都差。像素级方法学习下游任务的可区分像素级表征，但缺乏 LUSS 任务需要的足够的类别相关表征。在同一个类别内比较，大多数在 LUSS 任务中表现良好的方法在下游任务中都能取得更好的性能。因此，LUSS 和下游任务需要不同的表征，但都会从高质量的表征中受益。我们证明了我们提出的 P2P 对齐和 D2S 监督在 LUSS 任务 (表7(a)) 和下游任务 (表7(b)) 的有效性。这两种策略都提高了 LUSS 和下游任务的性能，表明了本文所提出的表征学习方法的通用性。

5.5 LUSS vs. WSSS

使用图像级标签的弱监督语义分割（WSSS）旨在仅用图像级标签实现对物体的分割语义。我们分析了 WSSS 方法的一些典型设定在 ImageNet-S_{50} 数据集上的影响，例如有监督 $ImageNet_{1k}$ 预训练模型 [37], [99], [100], [101], [103], 图像级 GT 标签 [101], [102], 和大型网络架构 [36], [37], [102], 并且我们展示了这些典型设置阻碍了将 WSSS 方法用于 LUSS 任务。除非另有说明，否则 WSSS 方法的设置与其官方设置保持一致。在无监督设置中，如果 GT 标签不可用，则使用自行生成的图像级伪标签替换 WSSS 方法中的 GT 标签。

预训练模型。 LUSS 的主要挑战之一是在没有监督的情况下学习有效的表征。然而，在 WSSS 方法中，对表征学习的影响研究较少，例如使用不同方法得到的预训练权重对 WSSS 方法的影像。现有的 WSSS 方法大多利用有监督的 $ImageNet_{1k}$ 预训练模型在例如 PASCAL VOC [10] 等语义分割数据集上微调模型 [37], [99], [100], [101], [103]。为了理解预训练的重要性，我们为 SEAM [36], SC-CAM [37], 和 AdvCAM [38] 使用不同的预训练模型，见表12。我们观察到，将 SEAM [36] 中的 $ImageNet_{1k}$ 预训练模型替换为在 $ImageNet_{50}$ 数据集上预训练的模型将 test mIoU 从 44.5% 降低到 35.8%。用 MoCo 和 SwAV 这两个无监督模型替换有监督模型，进一步将 test mIoU 分别降低到 19.1% 和 22.3%。SC-CAM 和 AdvCAM 都

表 12. WSSS 方法到 LUSS 任务的消融实验。WSSS 中的属性，例如监督预训练的模型、图像级 GT 标签和大型网络，不适用于 LUSS，这使 WSSS 方法在 LUSS 任务中的性能大幅下降。

ImageNet-S_{50}	Arch.	Param./MACC	Pre-train	Labels	mIoU		Img-Acc		F_β	
					val	test	val	test	val	test
SEAM [36]	ResNet-38 [122]	105.5M/100.4G	Sup. ImageNet$_{1k}$	GT	49.7	49.6	96.6	95.7	61.5	60.9
	ResNet-18 [108]	11.3M/1.9G	Sup. ImageNet$_{1k}$	GT	45.2	44.5	90.9	90.4	55.9	54.5
	ResNet-18 [108]	11.3M/1.9G	Sup. ImageNet-50	GT	35.1	35.8	81.2	81.5	46.3	46.5
	ResNet-18 [108]	11.3M/1.9G	MoCo. ImageNet-S_{50}	-	19.0	19.1	45.1	46.7	45.1	45.3
	ResNet-18 [108]	11.3M/1.9G	SwAV. ImageNet-S_{50}	-	22.1	22.3	54.6	53.5	41.1	41.1
SC-CAM [37]	ResNet-18 [108]	11.5M/1.8G	Sup. ImageNet$_{1k}$	GT	38.5	39.3	81.9	83.8	49.4	49.6
	ResNet-18 [108]	11.5M/1.8G	Sup. ImageNet$_{50}$	GT	31.3	32.1	70.2	71.0	44.1	44.4
	ResNet-18 [108]	11.5M/1.8G	MoCo. ImageNet-S_{50}	-	17.7	18.1	43.7	45.7	39.7	40.0
	ResNet-18 [108]	11.5M/1.8G	SwAV. ImageNet-S_{50}	-	19.0	19.7	50.0	49.1	38.6	40.8
SEAM [36] +AdvCAM [38]	ResNet-18 [108]	11.3M/1.9G	Sup. ImageNet$_{1k}$	GT	46.9	46.2	90.9	90.4	58.4	57.5
	ResNet-18 [108]	11.3M/1.9G	Sup. ImageNet$_{50}$	GT	36.9	37.6	81.2	81.5	49.2	49.6
	ResNet-18 [108]	11.3M/1.9G	MoCo. ImageNet-S_{50}	-	19.2	19.5	45.1	46.7	46.8	47.3
	ResNet-18 [108]	11.3M/1.9G	SwAV. ImageNet-S_{50}	-	23.7	23.3	54.6	53.5	44.2	43.9

表 13. 使用 ImageNet-S_{50}/ImageNet-S 数据集的半监督语义分割（半监督评测协议）。我们的 $_{s/p}$ 分别代表使用 SwAV [33] 和 PixelPro [35] 作为(5)中的 L_e。Supervised 是指使用图像级监督预训练初始化的模型。

Semi-supervised	mIoU		Img-Acc		F_β	
	val	test	val	test	val	test
ImageNet-S_{300}						
Supervised	27.7	27.5	61.1	62.3	64.3	64.9
SimCLR [26]	12.7	12.6	34.4	34.8	59.1	59.6
BYOL [34]	10.5	10.6	30.1	30.5	58.5	59.0
MoCov2 [25], [120]	12.6	12.3	33.0	32.5	59.2	59.4
DenseCL [95]	16.2	16.0	34.9	35.7	61.0	60.9
AdCo [121]	19.6	19.6	45.4	45.4	63.8	63.8
PCL [93]	17.3	17.4	41.7	41.8	61.7	61.9
SwAV [33]	23.0	23.3	51.2	51.5	64.0	64.0
PASS$_s$	25.7	25.7	52.3	52.8	65.5	66.0
PixelPro [35]	23.3	23.4	49.0	48.9	66.0	66.6
PASS$_p$	29.7	29.8	56.9	56.9	68.1	68.5
ImageNet-S						
Supervised	25.7	25.0	57.3	57.4	66.3	66.7
PixelPro [35]	16.0	15.6	36.0	36.2	66.2	66.5
PASS$_p$	18.9	18.6	40.9	41.3	68.0	68.4
SwAV [33]	18.2	17.9	42.8	43.2	66.0	66.2
PASS$_s$	19.4	19.2	43.3	43.4	66.6	66.9

遇到了相同的问题，这表明 WSSS 方法严重依赖于有监督的预训练模型。缺乏有监督的预训练使得表征学习对 LUSS 任务至关重要。并且我们的 ImageNet-S 数据集为公平评价预训练模型的表征质量提供了一个基础。

图像级 GT 标签。 WSSS 和 LUSS 任务之间的一个基本区别是 WSSS 需要图像级 GT 标签。类激活映射 (CAM) [125], [126] 通常被视为初始段区域，通常覆盖目标物体中最具区分度的小部分区域。许多 WSSS 方法严重依赖 GT 标签将 CAM 区域扩展到整个物体，并通过图像擦除 [98], [127], [128], [129], 区域增长 [97], [130], [131], [132], [133], 随机特征选择 [134], [135], 梯度操纵 [38], 或者数据集级信息 [136] 来更正错误的区域 [137]。为了分析 GT 标签对 WSSS 方法的影响，我们将最近的工作 AdvCAM [38] 应用于 SEAM [36]。AdvCAM 通过根据 GT 标签沿像素梯度扰动图像来反对抗地优化 CAM 结果。表12展示了使用 GT 标签的 AdvCAM 在基于 ImageNet$_{1k}$ 和 ImageNet$_{50}$ 有监督预训练模型的方法上在 test mIoU 上分别提升了 1.7% 和 1.8%。然而，当使用生成的伪标签和 MoCo 预训练模型时，性能增益仅为 0.4%。使用具有更好图像级精度的 SwAV 预训练模型，AdvCAM 将模型性能提高 1.0%。类似地，使用 GT 标签的 SEAM 和 SC-CAM 的性能相较于无监督设定有很大的优势。因此，由于缺少图像级 GT 标签，GT 标签的依赖性使得 WSSS 方法无法直接应用到 LUSS 任务。

网络架构。 许多网络体系结构被提出用来改进 WSSS，其中包括多尺度增强 [138] 和相关性预测 [99], [100], [103]。由于 PASCAL VOC 数据集的规模较小，许多最先进的 WSSS 方法使用具有大量参数和高计算成本的大型模型来提高性能，例如 wide ResNet-38 [36], [37], [102], [122] 和小输出步幅的 ResNet [38], [105]。由于本文提出的 ImageNet-S 数据集比 PASCAL VOC 大44 到800 倍，因此使用 WSSS 方法用的大型模型训练 LUSS 模型的计算成本非常高。为了分析模型架构的影响，我们更改了 [36] 中的网络（见表12）。为了公平比较，我们删除了 WSSS 方法中的 Deeplab 重新训练步骤。当使用标准 ResNet-18 [108] 替换 ResNet-38 [122] 时，test mIoU 从 49.6% 掉到 44.5%。大型模型有助于提高性能，但高计算成本使得无监督的 LUSS 模型训练变得不可行。

表 14. 在 ImageNet-S 测试集上使用距离匹配评测协议的有监督骨干模型的 mIoU 结果。Top-1 Acc. 是在 ImageNet-S 测试集上的分类准确率。* 表示模型使用 ImageNet-S 的训练集进行半监督微调训练。

Supervised	Top-1 Acc.	mIoU
ImageNet-S		
ResNet-50 [108]	83.6	29.8
ResNet-101 [108]	84.3	31.4
DenseNet-161 [139]	84.3	29.8
Inception V3 [140]	77.7	29.9
ResNeXt-50 [109]	84.4	32.6
ResNeXt-101 [109]	85.5	34.8
EfficientNet-B3 [141]	85.3	32.3
Res2Net-50 [110]	84.8	35.7
Res2Net-101 [110]	85.6	37.2
Swin-S [114]	87.8	38.6
Swin-B [114]	88.0	38.2
ConvNeXt-T* [142]	-	45.1
RF-ConvNeXt-T* (SingleRF) [143]	-	46.2
RF-ConvNeXt-T* (MultipleRF) [143]	-	47.0

5.6 ImageNet-S 数据集的应用

本文提出的 ImageNet-S 数据集有像素级标注，因此可以支持除 LUSS 任务以外的更多应用。本节介绍 ImageNet-S 数据集在大规模半监督语义分割、图像级有监督主干模型评测的应用和用于显著物体检测的子数据集。

大规模半监督语义分割。半监督语义分割需要使用一小部分标记数据和许多未标记数据进行训练。在 ImageNet-S 数据集的 1% 有像素级标注的训练图像上微调训练的 LUSS 模型，可实现半监督语义分割，这也是第3.2.1节中介绍的 LUSS 半监督评测协议。我们遵循在第5.1节中介绍的微调步骤的训练策略，唯一不同的是此处模型是使用 GT 标签用30 个迭代轮次训练的。半监督语义分割结果见表13。我们提出的方法优于 SwAV 和 PixelPro 基准，分别在 ImageNet-S_{300} 和 ImageNet-S 数据集上有可观的提升。我们基于 PixelPro 的方法甚至超过了 ImageNet-S_{300} 数据集上的图像级有监督模型。在半监督范式中，PixelPro 的性能与 SwAV 相似，但 SwAV 在距离匹配评测结果方面比 Pixelro 有很大优势 (见表9)。我们的结论是，具有像素级 GT 标签的微调模型使模型需要较少的自学习得到的类别相关表征能力。

评测有监督的主干模型。除了 LUSS 任务，ImageNet-S 数据集还可以评测经过图像级监督训练的主干模型的形状和类别表征能力。我们使用距离匹配评测协议在 ImageNet-S 测试集上对骨干模型的 mIoU 进行基准测试，见表14。作为参考，我们还提供了这些模型在 ImageNet-S 数据集上的 top-1 分类精度。我们观察到，图像级别的 top-1 精度并不总是与 mIoU 保持一致，这表明具有良好类别表征的模型可能不擅长形状表征。基于 Swin-transformer [114] 的自注意力在类别和形状表征上都表现很好。为了评测 ImageNet-S 数据集的性能在多大程度上受益于好的主干模型，我们测试了最近提出的 RF-ConvNeXt [143] 模型。该模型通过更合适的感受野组合增强了 ConvNeXt [142] 模型。RF-ConvNeXt 展现了高的语义分割性能，说明一个好的主干模型对 ImageNet-S 数据集非常有必要。

使用 ImageNet-S 子集进行显著性物体检测。显著性物体检测 (SOD) 旨在不考虑类别的情况下分割显著性物体 [117]。由于 SOD 的类别不敏感特性 [48]，无监督 SOD 模型可以为 LUSS 模型提供形状先验知识。为了促进大规模数据下的 SOD 任务，我们从 ImageNet-S 数据集中选择具有显著物体的图像，构建了一个 SOD 数据集，即 ImageNet-Sal。对于 ImageNet-S 训练/验证/测试集中的像素级标记图像，我们手动选择具有显著物体的图像，并删除非显著物体的标注。对于训练集中未标记的图像，我们借助于几个预先训练的 SOD 模型来选取具有显著物体的图像。由于选取的图像可能不包含显著物体，我们鼓励未来的 SOD 方法能够实现自我识别具有显著物体的训练图像。

6 结论

这项工作提出了一个新的大规模无监督语义分割问题，以便于在具有丰富多样性和大规模数据的现实环境中进行语义分割。我们为 LUSS 任务提供了一个基准，包含具有高度多样性的大规模数据、明确的任务目标和充分的评测方法。我们提出了一种新的 LUSS 方法，可以在没有人工标注监督的情况下，通过从大规模数据中学习类别和形状表征，并为像素分配标签。该 LUSS 方法包含增强的表征学习和像素注意力辅助的像素级标签生成策略。我们用多种评测方法评测了我们的方法，并揭示了 LUSS 对如语义分割等像素级下游任务的潜力。此外，我们对无监督表征学习方法和弱监督语义分割方法进行了评测和分析，总结了 LUSS 面临的挑战和可能的研究方向。

附录 A
基于 CAM 的推理

我们介绍了本文中使用的基于 CAM 的分割掩码推理策略。对于每张图像，我们有预测类别 c 的 CAM 图 $A^c \in R^{H\times W}$。然后，CAM 图被如下方式进行归一化：

$$\hat{A}^c = \frac{A^c - \min A^c}{\max A^c - \min A^c}. \tag{9}$$

然后，我们基于激活值将标签 c 分配给对应区域。具体来讲，在位置 (x, y) 分配的标签是

$$\mathrm{f}_{(\mathrm{x,y})} = \begin{cases} 0 & \hat{A}^k(x,y) < \tau; \\ k & \hat{A}^k(x,y) \geq \tau. \end{cases} \tag{10}$$

表 15. 使用完全无监督评测方法在 ImageNet-S 数据集上比较提出的 LUSS 基准和现存的 USS 方法。评测时不考虑“其他”类别。

LUSS	Pior	mIoU		b-mIoU		Img-Acc		F_β		mIoU				b-mIoU			
		val	test	val	test	val	test	val	test	S.	M.S.	M.L.	L.	S.	M.S.	M.L.	L.
ImageNet-S_{50}																	
MDC [46], [88]	-	4.0	3.7	1.4	1.2	14.9	13.4	31.6	31.3	0.4	2.6	3.8	5.0	0.2	1.1	1.4	1.5
MDC [46], [88]	I	14.9	14.6	3.2	3.1	44.8	40.8	33.2	32.6	2.7	11.1	14.93	19.51	0.9	2.2	3.2	4.8
PiCIE [46]	-	5.0	4.6	1.8	1.6	15.8	14.0	14.6	32.2	0.2	3.1	5.1	5.4	0.1	1.2	1.7	1.9
PiCIE [46]	I	18.1	17.9	3.7	4.0	45.0	44.0	32.1	31.6	4.4	13.3	20.4	23.5	0.9	2.7	4.4	5.8
MaskCon [31]	S	23.5	23.1	14.8	14.3	47.9	47.6	65.7	66.2	10.2	24.4	23.7	19.7	8.9	16.1	13.7	9.9
MaskCon [31]†	S	12.7	9.2	7.6	5.3	30.2	22.4	62.6	62.3	1.5	8.1	9.9	9.7	1.2	5.3	5.6	4.8
$PASS_s$	-	28.4	28.6	6.9	6.7	66.2	65.5	49.0	49.0	4.6	24.1	32.5	32.4	1.9	5.4	7.4	8.9
$PASS_p$	-	31.9	31.5	6.6	6.6	62.9	64.1	48.7	47.9	8.5	25.6	36.0	40.3	4.1	5.1	7.2	9.8
$PASS_p$+ RC	S	42.0	41.5	16.8	17.1	58.8	61.8	62.1	61.3	15.8	37.9	44.8	43.2	10.3	16.5	18.3	16.4
$PASS_p$+ Sal	S	42.5	41.5	19.7	19.5	64.6	65.2	70.0	69.9	17.2	40.8	44.4	37.8	13.6	21.8	19.9	14.7
ImageNet-S_{300}																	
$PASS_p$	-	16.6	16.0	4.4	4.2	34.7	32.8	34.4	34.3	2.8	12.0	16.5	21.8	1.4	3.2	3.9	6.4
$PASS_s$	-	17.9	18.0	5.1	5.1	43.9	42.6	47.6	47.5	4.0	13.4	19.4	23.5	1.9	4.1	5.4	7.0
ImageNet-S																	
$PASS_p$	-	7.3	6.6	2.4	2.1	19.9	18.0	34.8	34.6	1.3	4.6	7.1	8.4	0.6	1.5	2.1	2.8
$PASS_s$	-	11.5	11.0	3.8	3.5	24.0	22.3	37.1	36.9	2.4	8.3	11.9	13.4	1.2	3.0	3.7	4.3

其中 τ 表示一个阈值。

附录 B
没有“其他”类别的性能

通常预测“其他”类别比剩余的类别更容易，因为“其他”类涵盖了更大的图像区域。因此，包含“其他”类别可能会使模型更容易获得更高的 mIoU。我们在表15中评测了不考虑“其他”类别的性能。我们发现在 ImageNet-$S_{300/919}$ 数据集上有/无“其他”类别的 mIoU 差异大多小于 0.1%。在 ImageNet-S_{50} 数据集上，mIoU 大约有 1% 的差异。由于 mIoU 是通过对所有类别的 IoU 进行平均得到的，因此“其他”类别对 mIoU 的影响很小。

致谢 This work is funded by the National Key Research and Development Program of China Grant No.2018AAA0100400, NSFC (62225604), and the Fundamental Research Funds for the Central Universities (Nankai University, NO. 63223050). 感谢来自从 Imperfect Data Challenge [144] 的部分像素级标注。

参考文献

[1] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in IEEE Conf. Comput. Vis. Pattern Recog., 2015, pp. 3431–3440.

[2] F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” in Int. Conf. Learn. Represent., 2015.

[3] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in Int. Conf. Medical image computing and computer-assisted intervention, 2015, pp. 234–241.

[4] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 12, pp. 2481–2495, 2017.

[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 Trans. Pattern Anal. Mach. Intell., vol. 40, no. 4, pp. 834–848, 2017.

[6] M. Zhen, J. Wang, L. Zhou, S. Li, T. Shen, J. Shang, T. Fang, and L. Quan, “Joint semantic segmentation and boundary detection using iterative pyramid contexts,” in IEEE Conf. Comput. Vis. Pattern Recog., June 2020.

[7] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in IEEE Conf. Comput. Vis. Pattern Recog., 2016.

[8] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, “Semantic understanding of scenes through the ade20k dataset,” Int. J. Comput. Vis., 2018.

[9] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, “Bdd100k: A diverse driving dataset for heterogeneous multitask learning,” in IEEE Conf. Comput. Vis. Pattern Recog., 2020, pp. 2636–2645.

[10] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes challenge: A retrospective,” Int. J. Comput. Vis., vol. 111, no. 1, pp. 98–136, 2015.

[11] H. Caesar, J. Uijlings, and V. Ferrari, “Coco-stuff: Thing and

stuff classes in context," in IEEE Conf. Comput. Vis. Pattern Recog., 2018.

[12] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in Eur. Conf. Comput. Vis., 2018.

[13] M. Fan, S. Lai, J. Huang, X. Wei, Z. Chai, J. Luo, and X. Wei, "Rethinking bisenet for real-time semantic segmentation," in IEEE Conf. Comput. Vis. Pattern Recog., 2021.

[14] L. Wang, D. Li, Y. Zhu, L. Tian, and Y. Shan, "Cross-dataset collaborative learning for semantic segmentation," in IEEE Conf. Comput. Vis. Pattern Recog., 2021.

[15] L. Zhu, D. Ji, S. Zhu, W. Gan, W. Wu, and J. Yan, "Learning statistical texture for semantic segmentation," in IEEE Conf. Comput. Vis. Pattern Recog., 2021.

[16] S. Zhao, Y. Wang, Z. Yang, and D. Cai, "Region mutual information loss for semantic segmentation," in Adv. Neural Inform. Process. Syst., 2019.

[17] J. Liu, J. He, J. Zhang, J. S. Ren, and H. Li, "Efficientfcn: Holistically-guided decoding for semantic segmentation," in Eur. Conf. Comput. Vis., 2020.

[18] X. Li, X. Li, L. Zhang, G. Cheng, J. Shi, Z. Lin, S. Tan, and Y. Tong, "Improving semantic segmentation via decoupled body and edge supervision," in Eur. Conf. Comput. Vis., 2020.

[19] J. Liu, J. He, J. S. Ren, Y. Qiao, and H. Li, "Learning to predict context-adaptive convolution for semantic segmentation," in Eur. Conf. Comput. Vis., 2020.

[20] Y. Yuan, X. Chen, X. Chen, and J. Wang, "Segmentation transformer: Object-contextual representations for semantic segmentation," in Eur. Conf. Comput. Vis., 2021.

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein et al., "Imagenet large scale visual recognition challenge," Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, 2015.

[22] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in Int. Conf. Comput. Vis., 2017, pp. 843–852.

[23] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. Van Der Maaten, "Exploring the limits of weakly supervised pretraining," in Eur. Conf. Comput. Vis., 2018, pp. 181–196.

[24] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in Eur. Conf. Comput. Vis., 2020.

[25] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in IEEE Conf. Comput. Vis. Pattern Recog., 2020.

[26] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in International Conference on Machine Learning (ICML), 2020, pp. 1597–1607.

[27] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in Int. Conf. Comput. Vis., 2017, pp. 2961–2969.

[28] Y. Ouali, C. Hudelot, and M. Tami, "Autoregressive unsupervised image segmentation," in Eur. Conf. Comput. Vis., 2020, pp. 142–158.

[29] X. Zhan, Z. Liu, P. Luo, X. Tang, and C. Loy, "Mix-and-match tuning for self-supervised semantic segmentation," in AAAI, vol. 32, 2018.

[30] J.-J. Hwang, S. X. Yu, J. Shi, M. D. Collins, T.-J. Yang, X. Zhang, and L.-C. Chen, "Segsort: Segmentation by discriminative sorting of segments," in Int. Conf. Comput. Vis., 2019, pp. 7334–7344.

[31] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, and L. Van Gool, "Unsupervised semantic segmentation by contrasting object mask proposals," in Int. Conf. Comput. Vis., 2021.

[32] L. Beyer, O. J. Hénaff, A. Kolesnikov, X. Zhai, and A. v. d. Oord, "Are we done with imagenet?" arXiv preprint arXiv:2006.07159, 2020.

[33] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in Adv. Neural Inform. Process. Syst., 2020.

[34] G. Jean-Bastien, S. Florian, A. Florent, T. Corentin, P. R. H., B. Elena, D. Carl, B. P. Avila, Z. G. Daniel, M. A. Gheshlaghi, P. Bilal, K. Koray, M. Rémi, and V. Michal, "Bootstrap your own latent - a new approach to self-supervised learning," Adv. Neural Inform. Process. Syst., 2020.

[35] Z. Xie, Y. Lin, Z. Zhang, Y. Cao, S. Lin, and H. Hu, "Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning," in IEEE Conf. Comput. Vis. Pattern Recog., 2021.

[36] 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., 2020.

[37] 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., 2020.

[38] 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., 2021.

[39] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing via label transfer," IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 12, pp. 2368–2382, 2011.

[40] B. C. Russell, A. Efros, J. Sivic, W. T. Freeman, and A. Zisserman, "Segmenting scenes by matching image composites," in Eur. Conf. Comput. Vis., 2009.

[41] J. Tighe and S. Lazebnik, "Superparsing: scalable nonparametric image parsing with superpixels," in Eur. Conf. Comput. Vis., 2010, pp. 352–365.

[42] T. Malisiewicz and A. A. Efros, "Recognition by association via learning per-exemplar distances," in IEEE Conf. Comput. Vis. Pattern Recog., 2008.

[43] D. R. Martin, C. C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," IEEE Trans. Pattern Anal. Mach. Intell., vol. 26, no. 5, pp. 530–549, 2004.

[44] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," Int. J. Comput. Vis., vol. 59, no. 2, pp. 167–181, 2004.

[45] X. Li, Y. Yang, Q. Zhao, T. Shen, Z. Lin, and H. Liu, "Spatial pyramid based graph reasoning for semantic segmentation," in IEEE Conf. Comput. Vis. Pattern Recog., June 2020.

[46] J. H. Cho, U. Mall, K. Bala, and B. Hariharan, "Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering," in IEEE Conf. Comput. Vis. Pattern Recog., 2021, pp. 16 794–16 804.

[47] S. Gao, Y.-Q. Tan, M.-M. Cheng, C. Lu, Y. Chen, and S. Yan, "Highly efficient salient object detection with 100k parameters," in Eur. Conf. Comput. Vis., 2020, pp. 702–721.

[48] M.-M. Cheng, S. Gao, A. Borji, Y.-Q. Tan, Z. Lin, and M. Wang, "A highly efficient model to study the semantics of salient object detection," IEEE Trans. Pattern Anal. Mach. Intell., 2021.

[49] S. Jenni and P. Favaro, "Self-supervised feature learning by learning to spot artifacts," in IEEE Conf. Comput. Vis. Pattern Recog., 2018, pp. 2733–2742.

[50] P. Bojanowski and A. Joulin, "Unsupervised learning by predicting noise," in International Conference on Machine Learning (ICML), 2017, pp. 517–526.

[51] L. Zhang, G.-J. Qi, L. Wang, and J. Luo, "Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data," in IEEE Conf. Comput. Vis. Pattern Recog., 2019, pp. 2547–2555.

[52] Y.-H. H. Tsai, Y. Wu, R. Salakhutdinov, and L.-P. Morency, "Self-supervised learning from a multi-view perspective," in Int. Conf. Learn. Represent., 2021.

[53] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in Eur. Conf. Comput. Vis., 2016, pp. 649–666.

[54] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification," ACM Trans. Graph., vol. 35, no. 4, pp. 1–11, 2016.

[55] G. Larsson, M. Maire, and G. Shakhnarovich, "Colorization as a proxy task for visual understanding," in IEEE Conf. Comput. Vis. Pattern Recog., 2017, pp. 6874–6883.

[56] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in Eur. Conf. Comput. Vis., 2016, pp. 69–84.

[57] M. Noroozi, A. Vinjimoor, P. Favaro, and H. Pirsiavash, "Boosting self-supervised learning via knowledge transfer," in IEEE Conf. Comput. Vis. Pattern Recog., 2018, pp. 9359–9367.

[58] I. Misra and L. v. d. Maaten, "Self-supervised learning of pretext-invariant representations," in IEEE Conf. Comput. Vis. Pattern Recog., 2020, pp. 6707–6717.

[59] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in IEEE Conf. Comput. Vis. Pattern Recog., 2016, pp. 2536–2544.

[60] J. Donahue, P. Krähenbühl, and T. Darrell, "Adversarial feature learning," in Int. Conf. Learn. Represent., 2017.

[61] J. Donahue and K. Simonyan, "Large scale adversarial representation learning," in Adv. Neural Inform. Process. Syst., 2019.

[62] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in Int. Conf. Comput. Vis., 2015, pp. 1422–1430.

[63] T. N. Mundhenk, D. Ho, and B. Y. Chen, "Improvements to context based self-supervised learning," in IEEE Conf. Comput. Vis. Pattern Recog., 2018, pp. 9339–9348.

[64] M. Noroozi, H. Pirsiavash, and P. Favaro, "Representation learning by learning to count," in Int. Conf. Comput. Vis., 2017, pp. 5898–5906.

[65] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in Int. Conf. Learn. Represent., 2018.

[66] Z. Ren and Y. J. Lee, "Cross-domain self-supervised multi-task feature learning using synthetic imagery," in IEEE Conf. Comput. Vis. Pattern Recog., 2018, pp. 762–771.

[67] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," arXiv preprint arXiv:1807.03748, 2018.

[68] O. Henaff, "Data-efficient image recognition with contrastive predictive coding," in International Conference on Machine Learning (ICML), 2020, pp. 4182–4192.

[69] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, "What makes for good views for contrastive learning," in Adv. Neural Inform. Process. Syst., 2020.

[70] X. Chen and K. He, "Exploring simple siamese representation learning," in IEEE Conf. Comput. Vis. Pattern Recog., 2021.

[71] A. YM., R. C., and V. A., "Self-labelling via simultaneous clustering and representation learning," in Int. Conf. Learn. Represent., 2020.

[72] J. Li, P. Zhou, C. Xiong, R. Socher, and S. C. Hoi, "Prototypical contrastive learning of unsupervised representations," Int. Conf. Learn. Represent., 2021.

[73] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," in Int. Conf. Learn. Represent., 2019.

[74] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," in Adv. Neural Inform. Process. Syst., 2019.

[75] M. Ye, X. Zhang, P. C. Yuen, and S.-F. Chang, "Unsupervised embedding learning via invariant and spreading instance feature," in IEEE Conf. Comput. Vis. Pattern Recog., 2019, pp. 6210–6219.

[76] Y. Cao, Z. Xie, B. Liu, Y. Lin, Z. Zhang, and H. Hu, "Parametric instance classification for unsupervised visual feature learning," in Adv. Neural Inform. Process. Syst., 2020.

[77] C.-Y. Chuang, J. Robinson, Y.-C. Lin, A. Torralba, and S. Jegelka, "Debiased contrastive learning," in Adv. Neural Inform. Process. Syst., H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33, 2020, pp. 8765–8775.

[78] F. Wang, H. Liu, D. Guo, and F. Sun, "Unsupervised representation learning by invariance propagation," in Adv. Neural Inform. Process. Syst., 2020.

[79] M. Patacchiola and A. Storkey, "Self-supervised relational reasoning for representation learning," in Adv. Neural Inform. Process. Syst., 2020.

[80] J. D. Robinson, C.-Y. Chuang, S. Sra, and S. Jegelka, "Contrastive learning with hard negative samples," in Int. Conf. Learn. Represent., 2021.

[81] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in IEEE Conf. Comput. Vis. Pattern Recog., 2005, pp. 539–546.

[82] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in IEEE Conf. Comput. Vis. Pattern Recog., 2006, pp. 1735–1742.

[83] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with convolutional neural networks," in Adv. Neural Inform. Process. Syst., 2014.

[84] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in IEEE Conf. Comput. Vis. Pattern Recog., 2018, pp. 3733–3742.

[85] Y. Tian, X. Chen, and S. Ganguli, "Understanding self-supervised learning dynamics without contrastive pairs," in International Conference on Machine Learning (ICML), 2021.

[86] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," arXiv preprint arXiv:2103.03230, 2021.

[87] C. Zhuang, A. L. Zhai, and D. Yamins, "Local aggregation for unsupervised learning of visual embeddings," in Int. Conf. Comput. Vis., 2019, pp. 6002–6012.

[88] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in Eur. Conf. Comput. Vis., 2018, pp. 132–149.

[89] M. Caron, P. Bojanowski, J. Mairal, and A. Joulin, "Unsupervised pre-training of image features on non-curated data," in Int. Conf. Comput. Vis., 2019, pp. 2959–2968.

[90] X. Ji, J. F. Henriques, and A. Vedaldi, "Invariant information clustering for unsupervised image classification and segmentation," in Int. Conf. Comput. Vis., 2019, pp. 9865–9874.

[91] X. Yan, I. Misra, A. Gupta, D. Ghadiyaram, and D. Mahajan, "Clusterfit: Improving generalization of visual representations," in IEEE Conf. Comput. Vis. Pattern Recog., 2020, pp. 6509–6518.

[92] X. Zhan, J. Xie, Z. Liu, Y.-S. Ong, and C. C. Loy, "Online deep clustering for unsupervised representation learning," in IEEE Conf. Comput. Vis. Pattern Recog., 2020.

[93] J. Li, P. Zhou, C. Xiong, and S. C. Hoi, "Prototypical contrastive learning of unsupervised representations," Int. Conf. Learn. Represent., 2021.

[94] B. Roh, W. Shin, I. Kim, and S. Kim, "Spatially consistent representation learning," in IEEE Conf. Comput. Vis. Pattern Recog., 2021.

[95] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li, "Dense contrastive learning for self-supervised visual pre-training," in IEEE Conf. Comput. Vis. Pattern Recog., 2021.

[96] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao, and S. Yan, "Stc: A simple to complex framework for weakly-supervised semantic segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 11, pp. 2314–2320, 2016.

[97] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in Eur. Conf. Comput. Vis., 2016, pp. 695–711.

[98] 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., 2018, pp. 9215–9223.

[99] J. Ahn, S. Cho, and S. Kwak, "Weakly supervised learning of instance segmentation with inter-pixel relations," in IEEE Conf. Comput. Vis. Pattern Recog., 2019, pp. 2209–2218.

[100] J. Fan, Z. Zhang, T. Tan, C. Song, and J. Xiao, "Cian: Cross-image affinity net for weakly supervised semantic segmentation," in AAAI, 2020.

[101] G. Sun, W. Wang, J. Dai, and L. Van Gool, "Mining cross-image semantics for weakly supervised semantic segmentation," in Eur. Conf. Comput. Vis., 2020.

[102] W. Shimoda and K. Yanai, "Self-supervised difference detection for weakly-supervised semantic segmentation," in Int. Conf. Comput. Vis., 2019, pp. 5208–5217.

[103] 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., 2018, pp. 4981–4990.

[104] 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., 2020.

[105] L. Chen, W. Wu, C. Fu, X. Han, and Y. Zhang, "Weakly supervised semantic segmentation with boundary exploration," in Eur. Conf. Comput. Vis., 2020, pp. 347–362.

[106] Y. Zeng, Y. Zhuge, H. Lu, and L. Zhang, "Joint learning of saliency detection and weakly supervised semantic segmentation," in Int. Conf. Comput. Vis., 2019.

[107] 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 Eur. Conf. Comput. Vis., 2014, pp. 740–755.

[108] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in IEEE Conf. Comput. Vis. Pattern Recog., 2016.

[109] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in IEEE Conf. Comput. Vis. Pattern Recog., 2017, pp. 5987–5995.

[110] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," IEEE TPAMI, vol. 43, no. 2, pp. 652–662, 2021.

[111] S. Gao, Q. Han, D. Li, P. Peng, M.-M. Cheng, and P. Peng, "Representative batch normalization with feature calibration," in IEEE Conf. Comput. Vis. Pattern Recog., 2021.

[112] E. Cole, X. Yang, K. Wilber, O. Mac Aodha, and S. Belongie, "When does contrastive visual representation learning work?" arXiv preprint arXiv:2105.05837, 2021.

[113] K. Kotar, G. Ilharco, L. Schmidt, K. Ehsani, and R. Mottaghi, "Contrasting contrastive self-supervised representation learning pipelines," in Int. Conf. Comput. Vis., 2021, pp. 9949–9959.

[114] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in Int. Conf. Comput. Vis., October 2021, pp. 10 012–10 022.

[115] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, T. Duerig, and V. Ferrari, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," Int. J. Comput. Vis., 2020.

[116] B. Cheng, R. Girshick, P. Dollár, A. C. Berg, and A. Kirillov, "Boundary IoU: Improving object-centric image segmentation evaluation," in IEEE Conf. Comput. Vis. Pattern Recog., 2021.

[117] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 3, pp. 569–582, 2015.

[118] N. Zhao, Z. Wu, R. W. H. Lau, and S. Lin, "What makes instance discrimination good for transfer learning?" in Int. Conf. Learn. Represent., 2021.

[119] A. Islam, C.-F. R. Chen, R. Panda, L. Karlinsky, R. Radke, and R. Feris, "A broad study on the transferability of visual representations with contrastive learning," in Int. Conf. Comput. Vis., October 2021, pp. 8845–8855.

[120] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," arXiv preprint arXiv:2003.04297, 2020.

[121] Q. Hu, X. Wang, W. Hu, and G.-J. Qi, "Adco: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries," in IEEE Conf. Comput. Vis. Pattern Recog., 2021.

[122] Z. Wu, C. Shen, and A. van den Hengel, "Wider or deeper: Revisiting the resnet model for visual recognition," Pattern Recognition, vol. 90, pp. 119–133, 2019.

[123] M. Everingham, L. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," Int. J. Comput. Vis., vol. 88, pp. 303–338, 2009.

[124] M. Contributors, "MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark," https://github.com/open-mmlab/mmsegmentation, 2020.

[125] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in IEEE Conf. Comput. Vis. Pattern Recog., 2016, pp. 2921–2929.

[126] 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 Int. Conf. Comput. Vis., 2017, pp. 618–626.

[127] K. K. Singh and Y. J. Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," in Int. Conf. Comput. Vis., 2017, pp. 3544–3553.

[128] 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., 2017, pp. 1568–1576.

[129] Q. Hou, P.-T. Jiang, Y. Wei, and M.-M. Cheng, "Self-erasing network for integral object attention," in Adv. Neural Inform. Process. Syst., 2018.

[130] 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., 2018, pp. 7014–7023.

[131] 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., 2018, pp. 1354–1362.

[132] 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., pp. 1–1, 2021.

[133] 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., 2022, pp. 16 886–16 896.

[134] 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., 2018, pp. 1325–1334.

[135] 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., 2019.

[136] Y. Liu, Y.-H. Wu, P. Wen, Y. 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., vol. 44, no. 3, pp. 1415–1428, 2021.

[137] G. Sun, S. Khan, W. Li, H. Cholakkal, F. Khan, and L. Van Gool, "Fixing localization errors to improve image classification," Eur. Conf. Comput. Vis., 2020.

[138] 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., 2018, pp. 7268–7277.

[139] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in IEEE Conf. Comput. Vis. Pattern Recog., 2017.

[140] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in IEEE Conf. Comput. Vis. Pattern Recog., 2016.

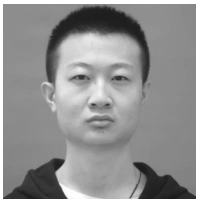
[141] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in International Conference on Machine Learning (ICML), 2019.

[142] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[143] S. Gao, Z.-Y. Li, Q. Han, M.-M. Cheng, and L. Wang, "Rf-next: Efficient receptive field search for convolutional neural networks," IEEE Trans. Pattern Anal. Mach. Intell., 2022.

[144] Y. Wei, "Learning from imperfect data (lid) challenge," https://lidchallenge.github.io/.

Shanghua Gao is a Ph.D. candidate in Media Computing Lab at Nankai University. He is supervised via Prof. Ming-Ming Cheng. His research interests include computer vision and representation learning.



Zhong-Yu Li is a Ph.D. student from the college of computer science, Nankai university. He is supervised via Prof. Ming-Ming cheng. His research interests include deep learning, machine learning and computer vision.

Ming-Hsuan Yang is a professor in Electrical Engineering and Computer Science at University of California, Merced. He received the PhD degree in Computer Science from the University of Illinois at Urbana-Champaign in 2000. Yang has served as an associate editor of the IEEE TPAMI, IJCV, CVIU, etc. He received the NSF CAREER award in 2012 and the Google Faculty Award in 2009.

Ming-Ming Cheng received his PhD degree from Tsinghua University in 2012, and then worked with Prof. Philip Torr in Oxford for 2 years. He is now a professor at Nankai University, leading the Media Computing Lab. His research interests includes computer vision and computer graphics. He received awards including ACM China Rising Star Award, IBM Global SUR Award, etc. He is a senior member of the IEEE and on the editorial boards of IEEE TPAMI and IEEE TIP.

Junwei Han is currently a Full Professor with Northwestern Polytechnical University, Xi′ an, China. His research interests include computer vision, multimedia processing, and brain imaging analysis. He is an Associate Editor of IEEE Trans. on Human-Machine Systems, Neurocomputing, Multidimensional Systems and Signal Processing, and Machine Vision and Applications.

Philip Torr received the PhD degree from Oxford University. After working for another three years at Oxford, he worked for six years for Microsoft Research, first in Redmond, then in Cambridge, founding the vision side of the Machine Learning and Perception Group. He is now a professor at Oxford University. He has won awards from top vision conferences, including ICCV, CVPR, ECCV, NIPS and BMVC. He is a senior member of the IEEE and Fellow of the Royal Society.