

## 摘要

本文展示了一个全新的数据集，目标是通过把物体识别放置在更广泛的场景理解问题之下，进而促进物体识别的发展。该数据集通过收集自然背景下复杂的日常生活场景图片而构建完成的。图片中的物体实例都进行了单独分割标注，这样做有助于提高物体定位的准确率。本数据集包含了91种物体类型的图像，这些物体类型能够被4岁大小的孩子毫不费力的识别出来。数据集有32.8万张图片，包含有250万个标注实例。众多众包人员通过使用先进的类别检测、实例标注、实例分割等用户界面创建了本数据集。本文提供了与PASCAL、ImageNet、SUN等数据集在统计学上的详细分析。最后，本文通过使用DPM提供边界框和分割检测结果作为基准性能分析。

## 1.引言

计算机视觉的一个首要任务之一就是对可见场景的理解。场景理解包含许多任务，包括识别呈现的是什么物体，2D和3D场景下物体定位，获取物体和场景的属性，物体特征之间的关系以及为场景提供语义描述。当前的物体分类和检测数据集[1][2][3][4]帮助我们探索场景理解的第一步。例如ImageNet数据集[1]，包含了前所未有的大规模的图片，使得物体分类和检测取得突破性的进展[5][6][7]。也有的机构创建了包含物体属性的数据集[8]，场景属性[9]，关键点[10]，和3D场景信息[11]。这就引出了一个显而易见的问题：哪一个数据集最适合继续促使我们向着场景理解的最终目标前进？

我们介绍一个大规模数据集，专注于场景理解中的三个核心问题：检测非图标视图（或者非标准布景[12]）中的物体，物体之间的上下文推理，物体精确的2D定位。对于目录中的很多物体类别都有图标。比如在基于网络的图片搜索中搜索物体类别‘自行车’，排名最靠前的结果往往是一张轮廓简洁的、中间没有遮挡的照片。我们假设当前的识别系统在图标类型的视图上表现优异，但是很难识别出日常生活中天天需要遇到的场景中的物体，而这些场景中物体往往位于背景中的、存在部分遮挡、姿态等杂乱的[13]。我们实验性的验证过模型在我们提供

的数据集上训练的结果要好于之前的那些数据集。在自然图片中识别出多个物体也是一种挑战。因为图片中物体存在小尺寸或者容易混淆的外观的问题，多个物体的分辨需要引入上下文。为了推进上下文推理的研究，需要用图像去描绘场景[3]而不是将物体之间孤立开。最后，我们详细的讨论了物体分布的空间理解将是场景分析中的核心部分。物体的空间分布能够粗略的使用边界框来定义[2]或者使用像素级别的精确分割[14][15][16]。正如我们示例的一样，为了测定这两种定义方式的性能，数据集的基本要求就是标注每一个物体的类别和全分割。我们的数据集特殊在于基于物体实例级别的分割，如图1所示。

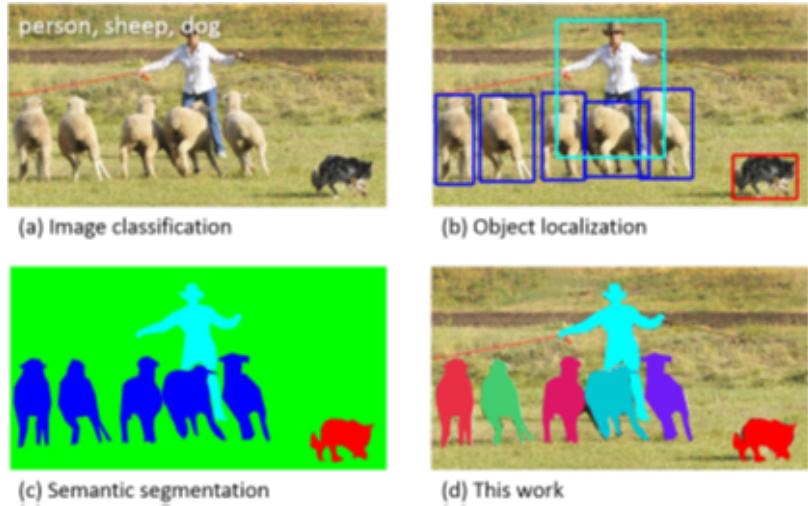


图1：当前许多物体识别数据集都关注于(a)图片分类，(b)物体边界框定位或者(c)像素级语义分割，我们的重点在于(d)单个物体实例的分割。我们引入了一个大的、有丰富标注的数据集，这些图片由日常生活中的生活场景图片组成。

为了构建实现上述三个目标的大规模数据集，我们使用了一种新颖的方法用于收集数据，大量的使用 Amazon Mechanical Turk服务（Amazon的众包服务）。第一也是最重要的，我们需要收集了一个大的数据集，包含了上下文关系和非标准的物体视图。在构建的过程中我们使用了一种非常简单的技术但是却很有效的方法，通过使用物体的名称对来查询图片从而将基于场景的图片查询关联在一起[17][3]。然后，使用多层次标注方法[18]对图片中包含的物体类别进行标注。对于每一个发现的类别，单独的实例将进行标注，验证，最后进行分割。考虑到标签的不确定，每一步的多方面权衡我们都将进行详细的探索。

MS COO (The Microsoft Common Objects in COntext )数据集包含有91中普遍的物体类别，其中82种拥有超过5K的标注实例，见Fig. 6。数据集总共拥有

32.8万张图片，包含250万个标注实例。对于著名的ImageNet数据集[1]，COCO拥有更少的物体类别但是每一个类别拥有更多的实例。这有助于具体的物体识别模型学习到更精确的2D定位能力。本数据集每个类别的实例数目也明显比PASCAL VOC[2]和SUN[3]数据集多。此外，我们的数据集跟别的数据集的一个明显区别就是每一张图片的标注实例数目有助于学习到上下文信息，Fig. 5。对比ImageNet(3.0)和PASCAL(2.3)，MS COCO每一张图片包含更多的物体实例(7.7)。和SUN数据集相比，SUN虽然拥有有意义的上下文信息，且每一张图片拥有超过17个物体和物体的填充，但是总体来说物体的实例数目更少。MS COCO这项工作的删减版本参考[19]。

## 2.相关工作

尽管计算机视觉的研究数据集的历史充满着批判角色。数据集不仅需要提供方法用于训练和评估算法，还需要促进研究走向更新更有挑战性的方向发展。如泉涌般创建的立体的和光学的真实数据集[20][21]极大刺激了人们对计算机视觉领域的广泛兴趣。早期的物体识别数据集的演进[22][23][24]帮助了上百种图像识别算法直接比较，与此同时也促进了这个领域研究更复杂的问题。最近，包含有百万图片的ImageNet数据集[1]在物体分类和检测的方面的研究因使用了深度学习算法取得了突破性成果[5][6][7]。物体识别相关的数据集能够粗略的根据专注的问题不同分为三个类别：物体分类、物体检测、场景语义标注。接下来将一个个介绍。

### 图片分类

物体分类任务需要二元标签注明物体是否在图片中出现；参考Fig. 1(a)。早期的数据集中的图片背景空白只有单个的物体出现，比如MNIST手写数字数据集[25]或者COIL日常物体数据集[26]。Caltech 101[22] 和Caltech 256[23]标志着开始转向从网络上检索的真实世界的物体，这种方式相应的将物体的类别从101扩展到256。流行机器学习社区中使用的数据集拥有很大数目的训练样本，CIFAR-10和CIFAR-100[27]分别提供了10种和100类别，图片分辨率都为32x32[28]。这些数据库包含有60k以上的图片以及上百种类别，但也只是我们真实世界中的一个很小的子集。最近，ImageNet[1]在数据集大小方面有了惊人的

增长。他们计划的数据集拥有22k个类别，每个类别有500–1000张图片。跟以往的数据集不同的是，以往的数据集只有大类[29]，比如狗或者单车[28]，ImageNet使用WordNet的层级[30]用于对大类进行细分成小类[31]。当前，ImageNet数据集包含有1400w张标注图片，并且已经明显的促进了图像分类的发展[5][6][7]。

### 物体检测

物体检测需要完成两个任务：一个是物体属于哪个类别，一个是物体在图片中的位置。物体的位置通常使用边界框来表示，Fig. 1 (b)。早起算法聚焦于人脸检测[32]使用多种混合数据集。在之后，更多的现实以及挑战性的人脸检测数据集被创建[33]。另一个流行的挑战是行人检测，相应的对个数据集也被创建[24][4]。The Caltech Pedestrian Dataset[4]包含有35w用边界框标注的实例。

对于基本物体类别的检测，从2005到2012经过多年的努力致力于创建和维护的多个基准数据集已经得到广泛的应用。The PASCAL VOC[2]数据集包含有20个物体类别11k图片。超过27k的物体实例使用了边界框进行标注，其中有7000进行了详细的分割。最近进行的一个物体检测挑战使用了ImageNet中40w中图片200个物体类别[34]。令人深刻的是其中35w个物体使用了边界框进行标注。

因为某些物体的检测比如太阳镜、手机、椅子高度的依赖上下文信息，因而检测自然环境中包含这些物体的数据集显得尤为重要。在我们的数据集中尽量收集上下文信息丰富的图片。边界框的精度也限制了检测算法准确率的评估。我们建议使用全分割实例来获取更高的检测准确率。

### 语义场景标注

场景中的语义实体标注任务需要对每一个像素进行分类，比如天空，椅子，楼梯，街道等等。对于检测任务，单个物体实例不需要进行分割，Fig. 1 (c)。这使得单个实例的物体标注变得难以定义，比如草地，街道或者墙。室内[11]和室外[35]场景的数据集都有一些数据集也包含了深度信息[11]。和语义场景标注类似，我们的目标是测量像素级别的物体标注准确率。同时，我们也聚焦单个物体实例之间的区分，这需要对物体范围有一个坚实的理解。

SUN[3]数据集是一个新颖的数据集，它将物体检测和语义场景标注数据集的许多属性联系在一起。SUN包含了来自WordNet[30]字典中908个场景类别，每个场景中的物体都进行了分割。3819个物体类别跨越了物体检测数据集的类别（人，椅子，猫）和语义场景标注（墙，天空，楼梯）。因而这个数据集通过收集图片描绘了多个场景类型，每一个物体类别的实例数目展示了长尾效应。也就是说，小部分类别拥有更大数目的实例（wall: 20,213, window: 16,080, chair: 7,971），然而更多的类别拥有极小的实例（boat: 349, airplane: 179, floor lamp: 276）。在我们的数据集中，我们可以确定每一个物体类别都有显著的实例数目，Fig. 5。

### 其他视觉数据集

这些数据集刺激了计算机视觉在多个领域的发展。一些重要的数据集包括Middlebury的立体[20]、多视图立体[36]和光流视觉[21]数据集。Berkeley Segmentation Data Set (BSDS500) [37]已经扩展用于评估分割和边缘检测算法。也有数据集用于识别场景[9]和物体属性[8][38]。大多数视觉领域都从这些竞赛的数据集获益取的长足的进步。



Fig. 2: Example of (a) iconic object images, (b) iconic scene images, and (c) non-iconic images.

## 3.图片收集

接下来将介绍如何选择物体的类别和候选图片。

### 3.1 普通物体类别

物体类别的选取是很重要的实践。类别的组织需要具有一定的代表性，和具体的应用相关以及具有一定的出现频率已支持大型数据集的收集。其他重要的决策就是是否包含‘物体’或者‘填充物’（填充物指的是天空、街道一类不好区分边界的东西）两种类别[39]，是否微调[31][1]以及物体一部分的类别是否应

该包含。“物体”指的是单个的物体实例，容易被区分和标注（人，椅子，车辆），“填充物”指的是那些没有明显边界的物质或者物体（天空、街道、草地）。既然我们的首要关注点是物体实例的准确定位，我们决定只包含“物体”类别而不是“填充物”。但“填充物”类别能提供明显的上下文信息，我们相信将来的对填充物进行标注是有好处的。

物体类别的特性是明显的。比如狗的类别可以使哺乳动物，狗，德国牧羊犬类别。为使得每一个类别都有显著数目的实例具备可实践性，我们选择限定数据集的上层类别，比如使用人们公认的类别标签（狗，椅子，人）。一些物体类别也可能是别的物体类别的一部分。比如脸是人身上的一个部分。我们相信包含对物体部分进行划分类别（脸，手）将对现实应用具有帮助。

我们使用多个数据源来建立顶层物体类别。我们首先联系PASCAL VOC[2]数据集的类别，还有1200种最经常见到的视觉可分辨的物体的一个子集[40]。为了进一步论证我们对物体类别的划分，我们组织了若干个4-8岁的儿童对他们看到的室内和室外的环境中物体进行命名。最终选定的272个类别在附录中给出。最后，研究者对每一个类别给出1-5分的得分，表明他们对该物体是否常见，是否对他们的实际的应用有用以及相对于其他类别的多样性。最后选出的票最高的类别，同时也能保证每一个子类别的数目均衡（动物，汽车，家具等）。类别中超过5000个实例也被移除。为了保证向后兼容，PASCAL VOC[2]中的类别都包含进来。最终选出的91个类别见Fig. 5(a)。

## 3.2 非图标图片收集

选定物体的类别之后，下一个目标就是收集相应的图片。我们粗略的将图片分为三类，Fig. 2: 图标类物体图片[41]，图标类场景图片[3]，非图标类图片。典型的图标类物体图片在图片中间有一个规范的大的物体影像，Fig. 2(a)。图标类场景是选择的是艺术的视角，通常场景没有人，Fig. 2(b)。图标类物体图片很容易就可以通过谷歌或者必应搜索得到。然而图标类场景图片能够获取到高质量的物体实例，但是往往缺乏上下文信息和正常视角。

我们的目标主要在于收集非图标图片，Fig. 2(c)。非图标图片包含的越多，泛化能力就越强[42]。我们在收集非图标图片时使用了两种策略。第一种是

PASCAL VOC[2]使用的方法，从Flickr中收集图片。Flickr包含了很多摄影爱好者上传的照片，这些照片都能够通过属性和关键词搜索到。第二种，我们不单单搜索物体的类别。搜索狗得到的是图标类图片。然而我们搜索成对的物体，比如狗+猫将得到更多的非图标类图片。意外的是，得到的这些图片不仅是包含这两种物体的，也包括大量的其他物体。对于我们用于搜索的物体类别对，参考附录。我们下载了单个摄影师在某个很短时间段拍的照片最多五张。在极少数情况下没有找到足够的图片，我们查询单个类别然后采用某个过滤策略将图标图片过滤掉。结果就是收到了32.8w张富有丰富上下文信息的图片，Fig2(c), Fig 6。

#### Annotation Pipeline

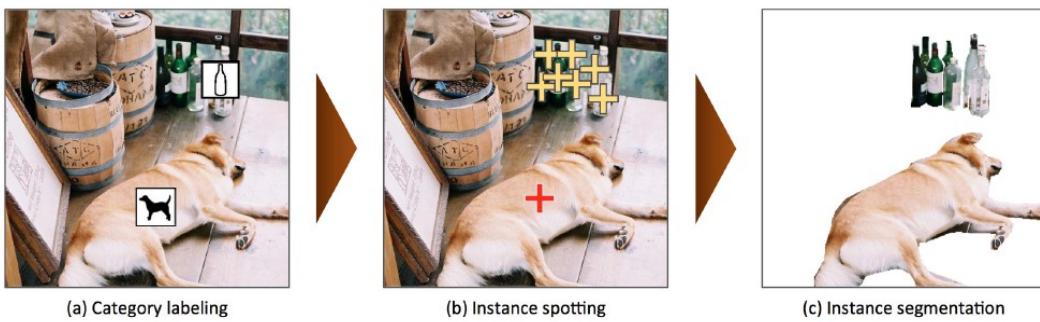


Fig. 3: Our annotation pipeline is split into 3 primary tasks: (a) labeling the categories present in the image (§4.1), (b) locating and marking all instances of the labeled categories (§4.2), and (c) segmenting each object instance (§4.3).

## 4.图片注解

接下来我们将介绍如何对收集到的图片进行注解。由于我们想要标注超过25亿个物体实例，高效和高质量的图片标注方法的设计尤为重要。注解方法如图Fig 3. 对于所有的群智任务，我们使用Amazon’s Mechanical Turk平台。用于接口的设计在附录中进行了描述。注意之前的版本[19]以来，我们已经使用了很多步骤促进注解的质量。尤其是我们已经将类别标注注解器和实例发现的阶段增加到8个。同时也增加一个阶段用于验证实例分割。

### 4.1类别标注

对数据集进行注解的第一步就是要判定每张图片中都有那些物体类别，Fig, 3(a)。因为我们有91个类别和大量的图片，要求标注人员对每张图片回答91个二分类问题成本会很高。取而代之的是，我们使用一种层次方法[18]。我们将物体类别归纳为11个大类别，参见附录。对于一张给定的图片，标注人员只需

要给出图片所属的大类别。比如说，标注人员可以很容易的判断照片中是否有狗，猫之类的物体出现在图片中。如果标注人员判定图片中出现了大类别的物体，比如动物，那么接着判断大类下面的小类，比如猫、狗，然后将类别的图片拖动到对应的实例上。这些图标的定位严格按照下面的步骤。我们强调，在本阶段每个类别有且之后一个实例需要被标注。为了保证高的召回率，8个标注人员对每一张图片进行标注。只有所有一个标注人员指明了这个类别，这个类别才被接纳。错误的标注将在接下来的阶段处理。详细的性能分析参见4.4节。这个阶段耗费了大约2w个工作时。

## 4.2 实例发现

在接下来的阶段里，每张图片中的实例都将被标注，Fig. 3 (b). 上一阶段，我们的标注人员为图片中的每个类别标注了一个实例，但是图片中往往包含多个实例。因此，我们安排了一个标注人员对上一阶段中存在多个实例的类别中每个实例使用十字架符号进行了标注。为了增加召回率，上一个阶段标注的物体实例的位置将展现给当前的标注人员。这样的做法帮助标注人员迅速的找到实例。标注人员还是用了放大镜对小实例进行标注。每一个标注人员被要求为每一张图片标注最多10个实例。这花费了1w个工作时。

## 4.3 实例分割

最后的阶段是最累的阶段，本阶段对每个实例进行分割，Fig. 3 (3)。在这个阶段，我们修改了Bell et al. [16] 的语义分割工具。我们使用该工具对上一阶段标注的物体实例进行分割。如果其他实例已经被分割了，那么这些实例将展现在标注人员面前。标注过程中有可能发现标注的物体类别没有实例，这可能是上一阶段的误标，或者已经全部标注好了。

分割250w个物体实例非常耗费时间，每1000个分割需要22个工作时。为了最小化花费，我们使用了一个标注人员对每一个实例进行标注。然而第一次完成标注的时候，物体的实例标注比较粗糙。所以，很重要的一点就是需要对标注人员进行标注培训。训练任务需要标注人员标注一个物体实例。标注人员需要将实例标注成跟真实的一样才算完成任务。培训提高了标注人员的标注准确率和分割结果，培训大概有三分之一的人通过。分割示例参考Fig. 6。培训淘汰了部分标

注质量差的标注人员，为保证每一个物体实例标注的结果正确我们对标注人员标注的结果进行了验证。有3-5个标注人员对标注的结果进行了检验。对标注不充分的实例将重新放入未标注集合进行重新标注。. 最后依然有一些标注人员的标注结果不好，我们将这部分标注结果移除了。

对于包含10个或者更少的物体实例的图片，每一个实例都单独分割，注意有些图片有多达15个实例也将分割。有些物体的实例相当的多，比如卡车上的香蕉，人群中的人。对于这些情况，将对所有的难以分辨单个实例的物体实例合在一起标注。在单个类别10-15个实例被标注之后，剩下的实例都标注为“群体”而都作为一体分割。为了方便评估，标注为“群体”的区域将被忽略，不影响检测算法的得分。

#### 4.4 注解性能分析

我们比较了普通标注人员和专业标注人员之间在类别标注任务上的标注质量，参考Fig. 4(a)。我们比较了7个专业标注人员和十个AMT标注人员之间的召回率和准确率。真实值使用了专业标注人员结果的多数。对于这个任务召回率很重要，因为错误的标注在接下来的阶段中将会被移除。Fig. 4(a) 展示8个AMT标注人员，和标签的数目相同，比专业标注人员得到更高的召回率。标注人员的召回率包含了9-10个AMT标注人员。物体类别的从属经常不清晰。Fig. 4(a) 中表明就算是专业的标注人员也对物体类别的归属意见不统一，比如因为图片固有歧义的原因或者对物体类别的划分不统一。假设清晰的标注有超过0.5的概率被标注，8个标注人员都失误的概率最多为 $0.5^8 \approx 0.004$ 。此外我们观察了增加注解器如何提升召回率，我们估计在实践中8个注解器检测到的超过99%的物体类别没有之后被作为假阳性被拒绝。注意，在实例发现中也是用了8个注解器，并且也有类似的结果。

最后，Fig. 4(b) 对AMT标注人员在更大数据集上类别标签的准确率和召回率进行了重新检查。圆圈大小表明标注人员的数目，圆圈颜色表示每一个标注人员的平均工作项目，圆圈位置表明准确率和召回率范围。不像Fig. 4(a)，我们使用留一法验证如果其他标注人员进行了类别标注之后，物体的类别是什么。因此，

大部分标注人员的标注准确率很高。完成最多工作的人拥有最高的准确率；所有在黑线下面的工作将被忽略。

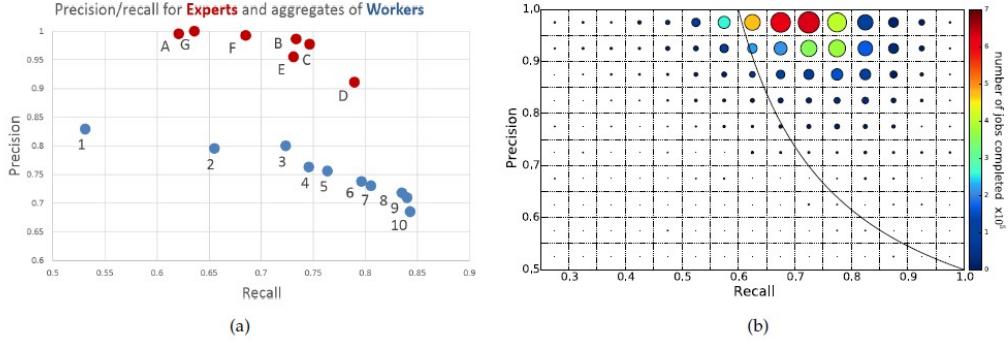


Fig. 4: Worker precision and recall for the category labeling task. (a) The union of multiple AMT workers (blue) has better recall than any expert (red). Ground truth was computed using majority vote of the experts. (b) Shows the number of workers (circle size) and average number of jobs per worker (circle color) for each precision/recall range. Most workers have high precision; such workers generally also complete more jobs. For this plot ground truth for each worker is the *union* of responses from all other AMT workers. See §4.4 for details.

## 4.5注解说明

我们为MS COCO数据集中每张图片添加了5个描述说明。完整的在统计上的描述，他们如何收集的将在另外的文章中简短介绍。

## 5.数据集统计信息

接下来我们将介绍MS COCO与其他流行数据集之间的不同之处，包括ImageNet[1]，PASCAL VOC 2012[2]，和 SUN[3]。每一个数据集在数据集的大小，类别的标注，图片的类型方面都有显著优点。ImageNet 收集了大量物体类别的图片，其中很多都有很好的处理。SUN专注于标注场景的类型以及该场景类型中的常见物体。PASCAL VOC的主要应用是自然图片中的物体检测。MS COCO的设计初衷适用于具有上下文信息的图片中的物体检测和分割。

91个类别中每个类别的实例数目参见Fig. 5(a)。数据集中每个物体类别图片数目以及每个类别实例数目参见Fig. 5(d)。虽然MS COCO对比ImageNet 和SUN具有更少的类别，但每一个类别拥有更多的实例，我们假定这能帮助复杂模型提高物体定位的准确率。对PASCAL VOC, MS COCO有更多的类别和实例。

我们的数据集一个重要的特点就是图片富含丰富上下文的真实图片。图片中的上下文信息可以通过图片中的物体类别和实例数目来评估，参考Fig. 5(b, c)。因为ImageNet中只有一个类别标签，我们只画出了验证集中的类别数目。我们的数据集中每一张图片平均拥有3.5个类别和7.7个实例。对比ImageNet 和

PASCAL VOC, 两者平均每张图片少于2个类别和3个实例。另一个有趣的发现, COCO中只有10%的图片拥有一个类别, 而ImageNet和PASCAL VOC中有60%。例外的是, SUN中拥有非常丰富的物体类别, 因为对物体类别没有进行限制。

最后我们分析了数据集中物体的平均尺寸。小物体更难分辨因而需要丰富的上下文信息。Fig. 5(e) 显示 MS COCO 和SUN中的物体尺寸都更小。

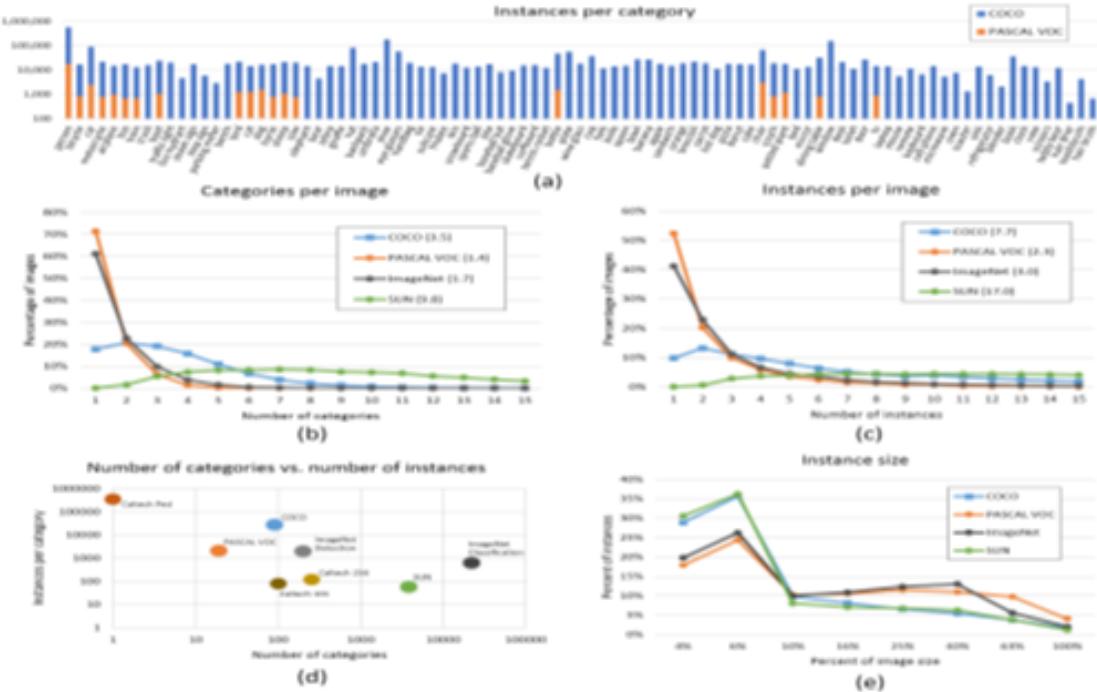


Fig. 5: (a) Number of annotated instances per category for MS COCO and PASCAL VOC. (b,c) Number of annotated categories and annotated instances, respectively, per image for MS COCO, ImageNet Detection, PASCAL VOC and SUN (average number of categories and instances are shown in parentheses). (d) Number of categories vs. the number of instances per category for a number of popular object recognition datasets. (e) The distribution of instance sizes for the MS COCO, ImageNet Detection, PASCAL VOC and SUN datasets.

## 6 数据集的划分

为了加快MS COCO数据集的发布, 我们将数据集分成了2个等同的子数据集。上半部分数据集于2014发布, 第二部分在2015年发布。2014版包含82,783训练图片, 40,504验证图片, 以及 40,775 测试图片(大约1/2训练, 1/4验证, 1/4测试)。2014训练和验证集合中有27w个人的分割实例以及 886k物体分割实例。累计到2015发布版数据集将包含165,482训练图片, 81,208验证图片, 和81,434测试图片。我们尽量使得数据集划分的时候近似图片的存在机会最小, 并显式的移除重复图片(使用[43]进行检测), 并且以摄影师的名称和时间进行分组。

根据已经建立的协议，训练集和验证集的数据将会发布，但是测试集不会。我们当前还在对用于测试集进行自动验证的评估服务器进行最后的工作。评估度量的完整讨论将在评估服务器发布之后进行。

注意2014发布版只有80个类别。剩下的11个类别没有进行分割：hat, shoe, eyeglasses（实例实在太多了），mirror, window, door, street sign（难以标注并且不容易区分），plate, desk（因为碗和桌子容易混淆）and blender, hair brush（实例极多）。这些类别中的某些将在2015版本中发布。

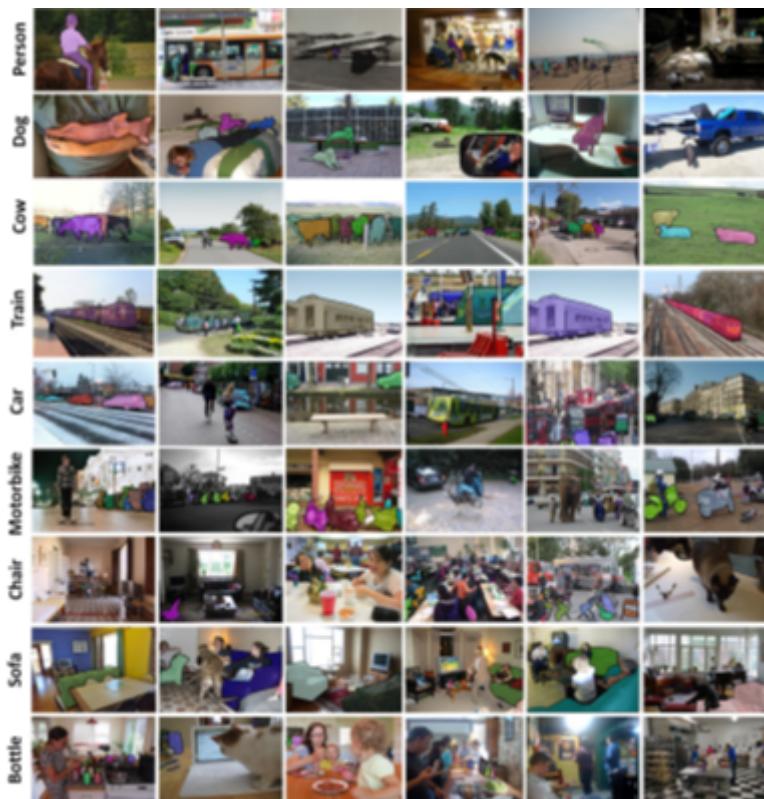


Fig. 6: Samples of annotated images in the MS COCO dataset.

## 7 算法分析

### 边界框检测

再接下来的实验中，我们从dataset1中选取了55000张图片并保留来自注解分割任务紧凑的边界框。模型的评估是在MS COCO和PASCAL数据集上同时进行的，参考Table 1。我们评估了2个不同的模型。DPMv5-P：最近的实现[44]（发布版本5[45]），在PASCAL VOC 2012数据集上训练。DPMv5-C：相同的实现，在COCO（5000正样本以及10000个负样本）。我们在COCO数据集上的模型使用相同的默认参数。

如果评估DPMv5-P在PASCAL VOC以及MS COCO数据集上的平均性能，可以发现在MS COCO上的表现要降低2个百分点，这表示MS COCO包含更丰富的图片，图片中的物体部分遮挡、杂乱无章等等。我们注意到在MS COCO (DPMv5-C) 数据集也有相似的性能下降。

通过比较DPMv5-P and DPMv5-C，我们将分析PASCAL VOC以及MS COCO数据集对检测性能的影响。两个模型使用相同的实现，不同的训练数据。表1表明DPMv5-C在PASCAL VOC数据集的20个类别中有6个类别优于DPMv5-P。在某些类别，比如狗、猫、人类，MS COCO数据集上的训练效果更差，但是其他类别比如公交车、电视和马上却表现的更好。

与过去的观察一致[46]，我们发现非标准场景中的图片对于，模型训练并总是能提升效果。这些图片有可能充当噪声，如果模型不能学习到这种外表的变种，已学习到的模型将会被破坏。我们的数据集考虑到了这方面的探索。

Torralba and Efros[42]建议提供一种标准用于衡量交叉数据集泛化能力，通过在一个数据集上训练在另一个数据集上测试的方式计算模型的性能丢失。通过这种方法DPMv5-P有12.7点准确率的差别 而DPMv5-C有7.7点准确率的差别。然而，在MS COCO上整体的性能较差。根据观察的结果，我们可以得出2个结论：1) MS COCO中的图片明显比PASCAL VOC中的复杂；2) 在复杂数据集MS COCO上训练的模型的泛化能力比稍简单数据集的PASCAL VOC强。对于两个数据集之间不同的更深入的探讨可以查看附录中对人以及椅子数据的可视化。

### **由检测生成分割**

现在我们将根据前人由物体检测生成物体分割的工作，描述一些简单的方法用于生成物体检测框和分割[47][48][49][50]。我们学习面向专家针对不同物体类别的像素等级的分割。通过将对齐训练实例中的分割掩码进行平均可以很轻松的学习到。我们的DPM检测器针对不同的混合模式学习到不同的掩码方式。

Fig. 7展示了相关示例。

### **通过分割评估检测**

就算物体检测能够提供正确的检测结果，但是分割依然需要对物体组成部分边界的良好定位，因而分割任务依然是一个具有挑战性的任务。为使得分割任务

的评估与检测无关，我们测试分割任务时只评估检测正确的部分。具体来说，就是只评估那些正确的预测了边界框的部分。那么实例分割预测部分与真实之间到底有多匹配呢？和物体检测的标准一样，我们标准是预测部分与真实部分需要重叠超过0.5，结果参考Fig. 8。

本数据集基准模型使用的是DPM。

Fig. 9 展示了DPM在PASCAL 20个类别上的分割结果以及在我们的数据集上的测试结果。

	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	moto	person	plant	sheep	sofa	train	tv	avg.
DPMv5-P	45.6	49.0	11.0	11.6	27.2	50.5	43.1	23.6	17.2	23.2	10.7	20.5	42.5	44.5	41.3	8.7	29.0	18.7	40.0	34.5	29.6
DPMv5-C	43.7	50.1	11.8	2.4	21.4	60.1	35.6	16.0	11.4	24.8	5.3	9.4	44.5	41.0	35.8	6.3	28.3	13.3	38.8	36.2	26.8
DPMv5-P	35.1	17.9	3.7	2.3	7	45.4	18.3	8.6	6.3	17	4.8	5.8	35.3	25.4	17.5	4.1	14.5	9.6	31.7	27.9	16.9
DPMv5-C	36.9	20.2	5.7	3.5	6.6	50.3	16.1	12.8	4.5	19.0	9.6	4.0	38.2	29.9	15.9	6.7	13.8	10.4	39.2	37.9	19.1

TABLE 1: Top: Detection performance evaluated on PASCAL VOC 2012. DPMv5-P is the performance reported by Girshick et al. in VOC release 5. DPMv5-C uses the same implementation, but is trained with MS COCO. Bottom: Performance evaluated on MS COCO for DPM models trained with PASCAL VOC 2012 (DPMv5-P) and MS COCO (DPMv5-C). For DPMv5-C we used 5000 positive and 10000 negative training examples. While MS COCO is considerably more challenging than PASCAL, use of more training data coupled with more sophisticated approaches [5], [6], [7] should improve performance substantially.

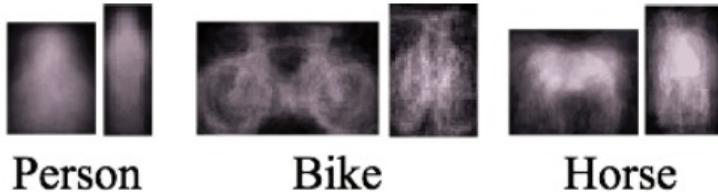


Fig. 7: We visualize our mixture-specific shape masks. We paste thresholded shape masks on each candidate detection to generate candidate segments.

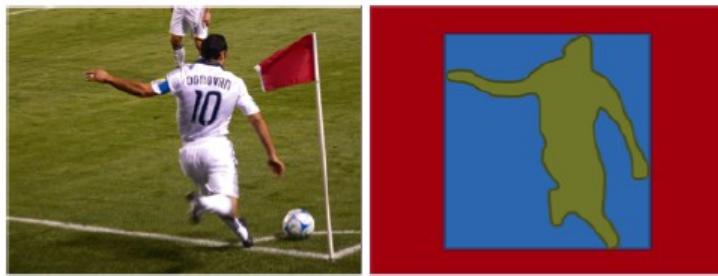


Fig. 8: Evaluating instance detections with segmentation masks versus bounding boxes. Bounding boxes are a particularly crude approximation for articulated objects; in this case, the majority of the pixels in the (blue) tight-fitting bounding-box do not lie on the object. Our (green) instance-level segmentation masks allows for a more accurate measure of object detection and localization.

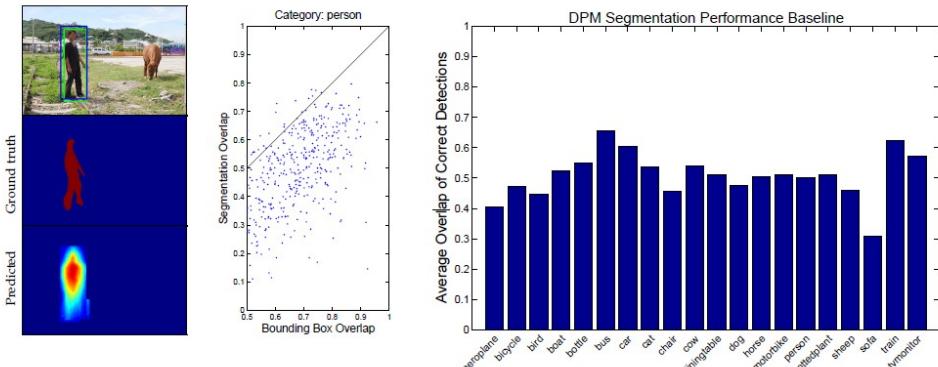


Fig. 9: A predicted segmentation might not recover object detail even though detection and ground truth bounding boxes overlap well (left). Sampling from the person category illustrates that predicting segmentations from top-down projection of DPM part masks is difficult even for correct detections (center). Average segmentation overlap measured on MS COCO for the 20 PASCAL VOC categories demonstrates the difficulty of the problem (right).

## 8 结论

本文介绍了一个用于物体检测和物体分割任务的，图片来自于自然生活场景中的新型数据集。

耗费了70,000个工作时，收集了广泛的物体数据，并对这些图片数据进行了标注以促进物体检测和物体分割算法的进步。图片的收集重点在与收集非标准场景以及多视图图片。数据集的统计结果表明图片中的物体之间包含了丰富的上下文信息。这为我们将来的标注工作提供了明确的方向。我们当前只标注了‘物体’，但是标注‘东西’能够提供更丰富的上下文信息并对检测任务能提供更多的帮助很多物体检测算法收益于额外的标注，比如实例标注的数目[4]或物体关键点的标注[10]。最后，我们的数据集为其他同类型的标注工作提供了一个好的参考基准，比如场景类型[3]，属性[9][8]和图片描述[51]。对于这些标注，我们也在积极的探索。下载和了解MS COCO数据集请参考我们的官网。MS COCO数据集将发展和增大；更新信息将在网上展示。

## 致谢

由Microsoft. P. P. and D. R. 资助的ONR MURI Grant N00014-10-1-0933提供资金用于众包人员聘请。最后感谢在数据集的构建和收集中提供过帮助的朋友和同事。

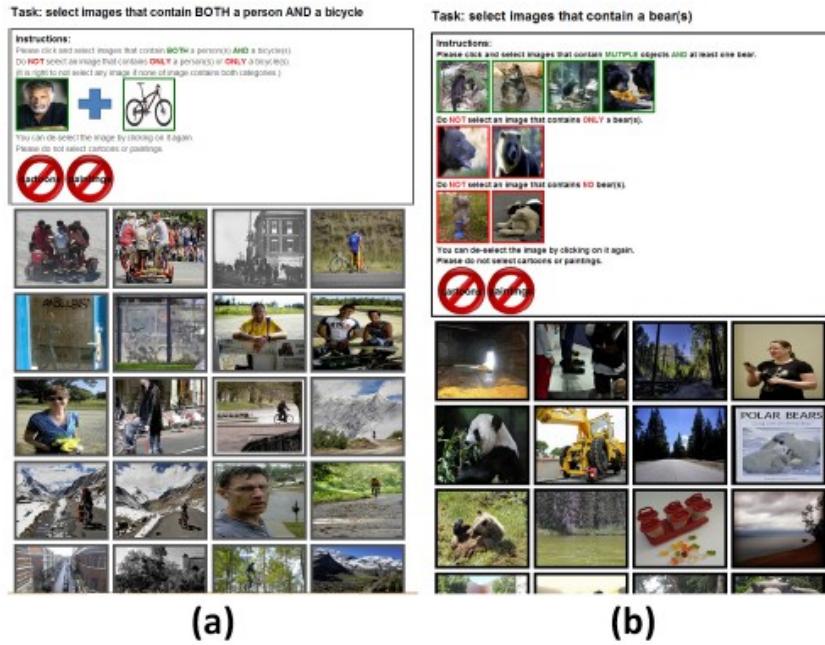


Fig. 10: User interfaces for non-iconic image collection. (a) Interface for selecting non-iconic images containing pairs of objects. (b) Interface for selecting non-iconic images for categories that rarely co-occurred with others.

## 附录

在附录中我们详细描述了AMT用户界面以及列出了所有272个候选类（我们从中只选择了91个类别）喝40个场景类别（用于场景中物体的索引）。

### 附录一：用户接口

我们描述并可视化用户界面，这些界面用于手机非标准场景图片，类别标注，实例标记，实例分割，分割验证以及最后的众包标注。

### 非标准图片收集

Flickr提供了丰富的关于图片的文字描述。但是图片有的不准确，并且质量不高。为了重新组织高质量的非标准图片描述，我们首先通过搜索物体对，在结果中选出候选图片，或者是物体、场景对。然后创建了一个AMT过滤任务允许用户从128张图片中移除非法的或者标准图片，Fig. 10。我们发现对于非标准图片和标准图片之间的区分标准很关键，因而提供了相应的示例图片。一些类别的物体与别的物体共现。这种情况下，我们只物体搜索阶段只搜索该物体，但也进行了相同的过滤操作，Fig. 10 (b)。

### 类别标注

Fig. 12(a) 展示类别标注的界面。我们鼓励众包人员尽量的标注出图片中的所有类别。众包人员通过拖动底部的类别图标到相应的物体实例上。图片中的一个类别的多个实例只需要标注一个。我们对类别按照父类进行分组，见图11，方便众包人员快速的跳过无关类别。

### 实例打点

Fig. 12(b) 描绘了标注所有给定类别实例的用户界面。界面由上一步类别标注结果初始化，此时使用十字叉标注。众包人员标注给定类别高达10个实例，在每一个实例上标注一个十字星。为了方便标注小的物体，我们提供了‘放大镜’将众包人员当前选中的区域放大两倍。

### 实例分割

Fig. 12(c) 展示的是实例分割界面。我们修改OpenSurfaces项目[16]以满足我们项目的研发，该项目定义了一个简单的AMT任务用于分割实际场景中多个区域中相同的材料。在我们的项目中，我们把上一步骤中标注的每一个单一物体当做分割的单位。为了加快分割的速度，我们添加了一个可视化的物体类别图标用于提醒众包人员分割物体的类别。关键的，我们也添加了放大功能以方便有效的标注小物体喝弯曲的边界。在上一步标注阶段，为了最大限度的覆盖所有的物体实例，我们对每一张图片使用了多个标注人员。但在本阶段对于一个物体实例可能存在不同的标注方式，因为为了消除这种不统一，我们将对实例分割任务队列化，已经标注过的实例将不被标注。

### 分割验证

Fig. 12(d) 展示了我们的分割验证界面。因为上一步标注任务消耗了太多的时间，本次标注只进行一次。验证阶段是为了保证上一步实例分割操作是正确的。标注人员从64个分割实例中选出质量较差的图片。64张中有4张是标注效果差的，标注人员必须从中选出3张才能完成本阶段工作。每一个分割实例首先交给3位标注人员进行分割，如果有任意一位表明这是质量差的分割，那么该图片将分发给另外2位标注人员进行重新分割。完成上述步骤之后如果没有获得超过4/5的人员认为是质量差的图片将认为是正确分割图片。否则将认为是未分割图片而打回。Fig. 15. 中展示了4/5认为是良好的图片则通过，3/5认为图片标注可以通过则返回未标注图片队列中重新进行标注。

## 群体标注

Fig. 12(e) 展示了行人标注界面。像我们之前的讨论，图片中有不超过10个实例则全部进行标注。但有些图片中实例的数量非常多。这种情况下群体标注将使用一种更高效的方法。我们允许标注人员对图片中实例所属像素进行‘绘制’，而不是使用多边形进行标注。群体标注和物体实例标注类似，但没有进行单个实例进行标注。我们确保只对需要标注的类别在实例超过10个的情况下才进行群体标注。

## 附录二：物体分类和场景分类

我们的数据集包含91个类别（2014年的发布版本包含其中80个类别的语义分割）。我们通过儿童从WordNet, labelMe, SUN以及其他数据集的类别中选出常见的物体类别。我们再从中选出相对有挑战性的272个类别，详情参考第三节。表2列出了全部的按照得票数降序排列的类别名称。

综上所述，最终选取了得票数最高的91个类别并保持父类（动物、车辆、家具等等）下的类别数目平衡。正如第三章讨论的一样，额外使用物体对进行检索对于收集非标准场景图片非常有效，物体-场景对同理。对于这个任务，我们选择了从SUN数据集中的40个场景类别选择了与物体类别经常共现的子集。表3列出了40个场景类别（20个室内，20个室外）。



Fig. 10: User interfaces for non-iconic image collection.  
(a) Interface for selecting non-iconic images containing pairs of objects. (b) Interface for selecting non-iconic images for categories that rarely co-occurred with others.

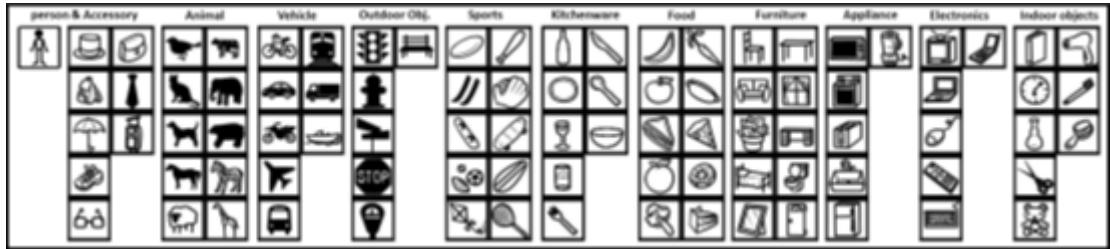


Fig. 11: Icons of 91 categories in the MS COCO dataset grouped by 11 super-categories. We use these icons in our annotation pipeline to help workers quickly reference the indicated object category.

(a) Category Labeling
(b) Instance Spotting
(c) Instance Segmentation

(d) Segmentation Verification
(e) Crowd Labeling

图12：用于收集实例标注的用户界面



(a) PASCAL VOC.

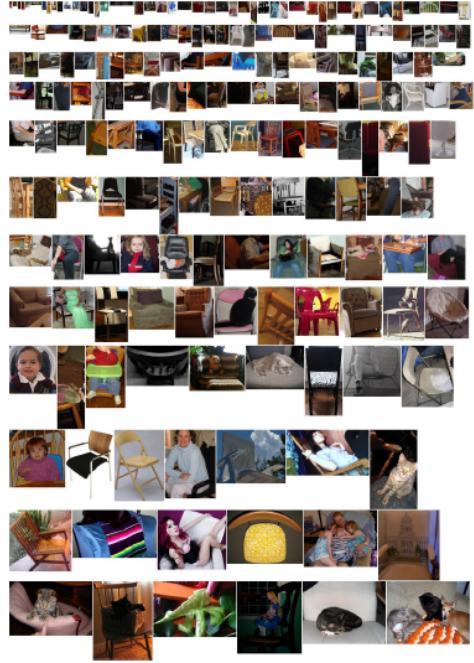


(b) MS COCO.

Fig. 13: Random person instances from PASCAL VOC and MS COCO. At most one instance is sampled per image.

person	bicycle	car	motorcycle	bird	cat	dog	horse	sheep	bottle
fridge	couch	potted plant	tv	cow	airplane	hat*	license plate	bed	laptop
banana	microwave	sink	oven	toaster	bus	train	mirror*	dining table	elephant
face	head	toilet	book	boat	plate*	cell phone	mouse	remote	clock
eye	hand	apple	keyboard	backpack	steering wheel	wine glass	chicken	zebra	shoe
street sign*	umbrella	scissors	rock	traffic light	eyeglasses*	cup	blender*	hair drier	wheel
headlights	window*	door*	fire hydrant	bowl	saucer	fork	knife	frisbee	bear
nose	teddy bear	desk*	computer	refrigerator	pizza	squirrel	duck	orange	guitar
printer	pan	tie	stop sign	surfboard	sandwich	pen/pencil	kite	chandelier	toothbrush
handbag	hot dog	head	sports ball	broccoli	suitcase	carrot	baseball bat	parking meter	fish
skis	stapler	table lamp	basketball hoop	donut	vase	chandeler	baseball glove	giraffe	jacket
skateboard	helicopter	tomato	egg	door handle	power outlet	tennis racket	tiger	table	coffee table
chopping board	washer	lion	tree	bunny	pillow	hair brush*	cake	feet	bench
goat	monitor	key	monkey	hair brush*	light switch	fan (on floor)	arms	house	cheese
ears	mouse	phone	picture frame	can	fan	fan (on floor)	legs	scarf	surf
playing cards	towel	hippo	strawberries	pumpkin	pillow	van	rabbit	owl	deer
tire	necklace	tablet	corn	dollar bill	light switch	kangaroo	rhinoceros	sailboat	muffins
toy cars	bracelet	bat	balloon	ladder	fan (on floor)	pineapple	meat	window	cookie
box	platypus	pancake	cabinet	gloves	fan (on floor)	milk	desktop	carpet	bacon
pasta	grapes	shark	swan	whale	dryer	dryer	camera	building	shorts
moon	road/street	fountain	fax machine	ingers	towel	torso	wheelchair	shirt	flip flops
basketball	telephone	movie (disc)	football	bat	hot air balloon	side table	lizard	gate	cabinets
radio	fence	goal net	goat	goose	long sleeve shirt	cereal	seahorse	raft	copier
shears	aardvark	dinosaur	toys	engine	soccer ball	field goal posts	rocket	rooster	seats
ipad	iphone	hoop	unicycle	honey	legos	fly	socks	tennis net	mat
goldfish	robot	crusher	hen	back	table cloth	soccer nets	roof	baseball	underpants
jetpack	robots	animal crackers	basketball court	horn	firefly	fireflies	turkey	pajamas	nectar

TABLE 2: Candidate category list (272). **Bold**: selected categories (91). **Bold\***: omitted categories in 2014 release (11).



(a) PASCAL VOC.



(b) MS COCO.

Fig. 14: Random chair instances from PASCAL VOC and MS COCO. At most one instance is sampled per image.

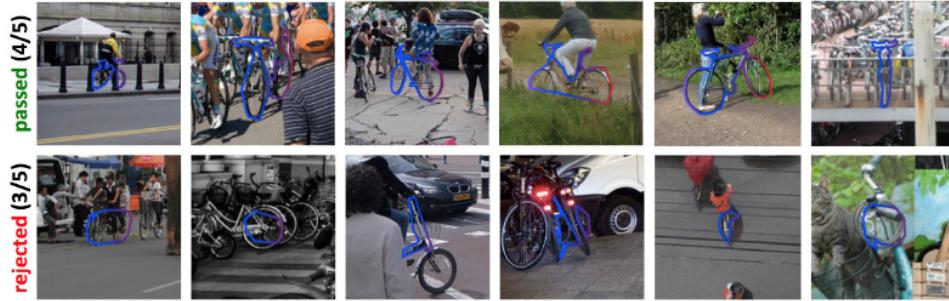


Fig. 15: Examples of borderline segmentations that passed (top) or were rejected (bottom) in the verification stage.

library	church	office	restaurant	kitchen	living room	bathroom	factory	campus	bedroom
child's room	dining room	auditorium	shop	home	hotel	classroom	cafeteria	hospital room	food court
street	park	beach	river	village	valley	market	harbor	yard	parking lot
lighthouse	railway	playground	swimming pool	forest	gas station	garden	farm	mountain	plaza

TABLE 3: Scene category list.